INDÍCE

1. Introdução

O presente texto tem como objectivo servir de apoio ao programa elaborado para a disciplina de Matemática para as Ciências Sociais.

Este texto não é um texto teórico, por onde os interessados poderão ir buscar os conhecimentos necessários para o estudo da disciplina, mas tão só um conjunto de exercícios que poderão esclarecer melhor o objectivo que tentámos imprimir à disciplina.

Assim, alguns dos exercícios propostos não apresentam as soluções, por pensarmos que são triviais. Efectivamente o que pretendemos não é apresentar exercícios complicados, mas antes pelo contrário, exercícios simples, mas variados, que sejam exemplos de assuntos tratados na realidade do dia a dia, sem que se pretenda, nos exemplos apresentados, esgotar este tema.

Quando pensamos que os temas propostos são susceptíveis de não estarem tão presentes nas pessoas a quem este texto se dirige, nomeadamente os Professores, aprofundamos um pouco mais o assunto. Um exemplo desta situação é o que se passa com o tema da Inferência Estatística.

2. Estatística

2.1. Interpretação de tabelas e gráficos através de exemplos.

Objectivos a atingir:

- ✓ Familiarizar os alunos com a leitura e interpretação de informação transmitida através de tabelas e gráficos.

De forma a cimentar alguns dos conhecimentos adquiridos no Ensino Básico, na introdução do tema Estatística, propomos que se comece com a interpretação de tabelas e gráficos, já construídos, que são instrumentos privilegiados em qualquer procedimento estatístico. Pretendemos chamar a atenção para o quanto estes processos podem ser ricos na transmissão de informação, mas também alertar para algumas representações que podem levar a interpretações erradas. Os exemplos devem ser sugestivos, ligados a actividades do mundo real.

Pretende-se que no fim deste módulo os alunos estejam familiarizados com os diferentes tipos de gráficos e tabelas, que são usados para reduzir a informação contida num conjunto de dados, sem terem a preocupação de quais as regras ou metodologias utilizadas na sua construção.

2.1.1. Exemplo 1 – Indicadores sobre a população continental e alentejana

Considere as seguintes tabelas que dão alguns indicadores genéricos, sociais e demográficos relativamente à população residente no Continente e à residente no Alentejo:

Indicadores Genéricos

Designação do indicador	Valor Contin.	Valor Alentejo	Unidade	Período
Área Total	88797.4	26931.2	Km^2	1997
Número de freguesias	4037	294	Nº	1997
Área Média das Freguesias	22	91.6	Km^2	1997
Densidade Populacional	106.5	19.1	hab/km^2	1997
Estimativa População Residente - Total	9454240	514790	Indivíduo	97/12/31
Estimativa População Residente - Homens	4553600	250030	Indivíduo.	1991
População Residente HM	9375926	543442	Indivíduo.	1991
Edifícios	2712866	234897	Nº	1991
Alojamentos Familiares Clássicos	3992163	267295	Nº	1991
Famílias Clássicas	3018089	193476	Nº	1991

Fonte: INE – Página INFOLINE – www.infoline.ine.pt

Indicadores Sociais

Designação do indicador	Valor Contin.	Valor Alentejo	Unidade	Período
Índice Per Capita do Poder de Compra (Portugal=100)	102	68	percentagem	1997
Médicos por 1000 Habitantes	3.1	1.4	Nº	1997
Camas Hospitalares por 1000 Habitantes	3.9	3.2	Nº	1997
Pensionistas activos por 1000 Habitantes	24	34.9	Nº	97/12/31
Pensão Média Anual por Pensionista Activo	424(1)	410(2)	Milhares de escudos	◊ 1996 ◊ 97/12/31
Alunos Matriculados no Sistema de Ensino	2025085	100188	Nº	1995/1996
Pessoal Docente do Ensino Público	141772	8397	Nº	1995/1996

Fonte: INE – Página INFOLINE – www.infoline.ine.pt

Indicadores Demográficos

Designação do indicador	Valor Contin.	Valor Alentejo	Unidade	Período
Taxa de Natalidade	11.3	9	permilagem	1997
Taxa de Mortalidade	10.5	14.6	permilagem	1997
Excedente de Vidas	0.7	-5.6	permilagem	1997
Taxa de Nupcialidade	6.6	5.3	permilagem	1997
Taxa de Divórcio	1.4	1	permilagem	1997
Índice de Envelhecimento	90.8	147.2	percentagem	1997/12/91

Fonte: INE – Página INFOLINE – www.infoline.ine.pt

Considere a tabela adequada para responder às seguintes questões:

a) O que significa Taxa de Natalidade? Como se calcula? Compare a taxa de natalidade no Continente e na região do Alentejo. O que conclui?

b) O que significa o termo permilagem? Faça a analogia com o termo percentagem.

c) Da consulta da tabela verifica-se que no Alentejo morrem mais pessoas do que nascem. Como é que se pode chegar a esta conclusão? O que é que pode concluir sobre o envelhecimento da população alentejana? Indique mais do que um indicador (de tabelas diferentes) que lhe permita tirar a conclusão que tirou.

d) Considera que a população alentejana tem um poder de compra idêntico ao resto do País? Explique a sua resposta.

e) Calcule uma estimativa da percentagem de indivíduos do sexo feminino residentes no Continente. Faça o mesmo para a região do Alentejo e compare os valores obtidos.

f) Diga se na sua opinião o Alentejo é altamente ou baixamente povoado. Em que é que se baseou para tirar essa conclusão?

2.1.2. Exemplo 2 – Estudo sobre a idade de veículos importados

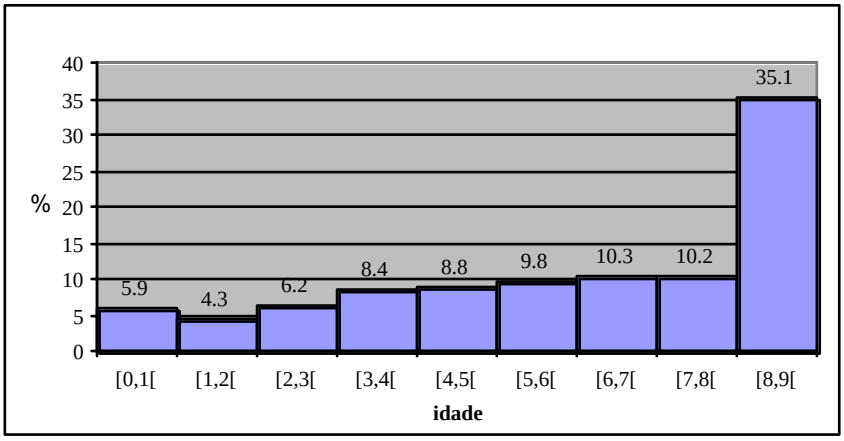
Considere a seguinte tabela que diz respeito à idade dos veículos usados introduzidos no consumo (importados):

Veículos automóveis	1994		1995		1996	
	Nº veic	%	Nº veic	%	Nº veic	%
< 1 ano de uso	198	3.28	806	4.49	1899	5.88
1 ano até 2 anos de uso	483	8.01	659	3.67	1389	4.30
2 anos até 3 anos de uso	368	6.10	751	4.18	1986	6.15
3 anos até 4 anos de uso	543	9.00	1255	6.99	2723	8.43
4 anos até 5 anos de uso	514	8.52	1461	8.13	2841	8.80
5 anos até 6 anos de uso	552	9.15	1701	9.47	3163	9.80
6 anos até 7 anos de uso	550	9.12	1810	10.08	3337	10.33
7 anos até 8 anos de uso	739	12.26	2076	11.56	3308	10.24
Com mais de 8 anos de uso	2063	34.21	7445	41.44	11645	36.06
Total	6030	100	17964	100	32291	100

Fonte: ANECRA – Revista nº 152

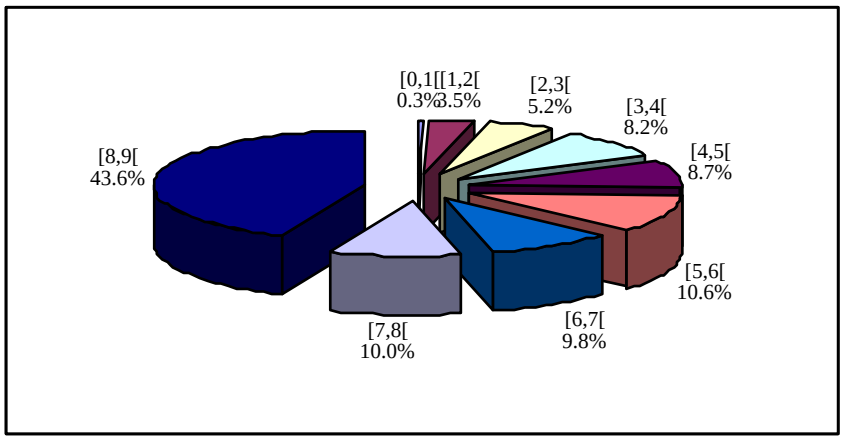
a) Da análise da tabela anterior o que é que conclui relativamente ao nº de veículos importados de 1994 a 1996? A que pensa que é devido esse facto?

b) Considere a seguinte representação gráfica – histograma, relativamente aos dados de 1996:



Qual o tipo de veículos que predomina? Considera a situação preocupante? Porquê?

c) Considere a seguinte representação gráfica que representa, para o ano de 1997, sob a forma de um diagrama circular, a distribuição por idades dos veículos ligeiros de passageiros, usados, introduzidos no consumo:



Fonte: ANECRA – Revista nº 152

Qual a percentagem de veículos ligeiros de passageiros importados com 7 ou mais anos de idade? Pensa que o nosso país está a receber a sucata da Europa, como é sugerido na revista da ANECRA?

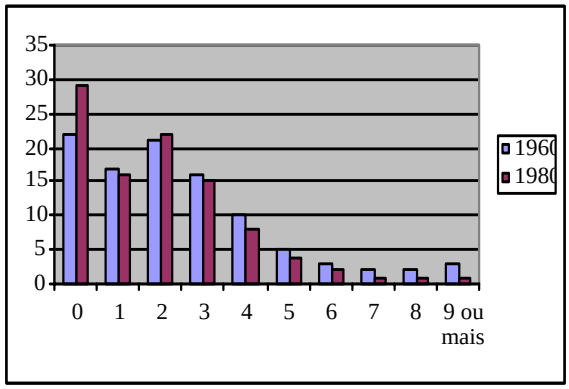
d) Suponha que nas representações gráficas anteriores não tinha indicado, associado à classe, a respectiva percentagem de veículos. Qual das representações gráficas considera mais elucidativa e que transmite de forma mais correcta a informação?

e) Estude a evolução de veículos importados nos anos considerados e faça um pequeno relatório comentando a situação (Refira a situação preocupante de Portugal ser um recordista europeu de acidentes e mortes na estrada).

2.1.2. Exemplo 3 – Número de filhos das famílias americanas

Considere a seguinte tabela de frequências e o correspondente diagrama de barras com informação respeitante ao nº de filhos das mulheres americanas com 18 ou mais anos de idade, relativamente a 1960 e a 1980 (Freedman, 1991):

Nºfilhos	1960	1980
0	22	29
1	17	16
2	21	22
3	16	15
4	10	8
5	5	4
6	3	2
7	2	1
8	2	1
9 ou mais	3	1



Faça um pequeno relatório comentando a situação, referindo nomeadamente implicações sociais.

2.1.3. Exemplo 4 (Thiessen, 1997) – Actividade económica por sector

Na tabela seguinte apresentam-se alguns dados, relativos à Alemanha, sobre a evolução da actividade económica por sector da sua economia:

Mudança estrutural da actividade económica da Alemanha, % partilhadas pela força de trabalho por sector de economia, 1882-1992

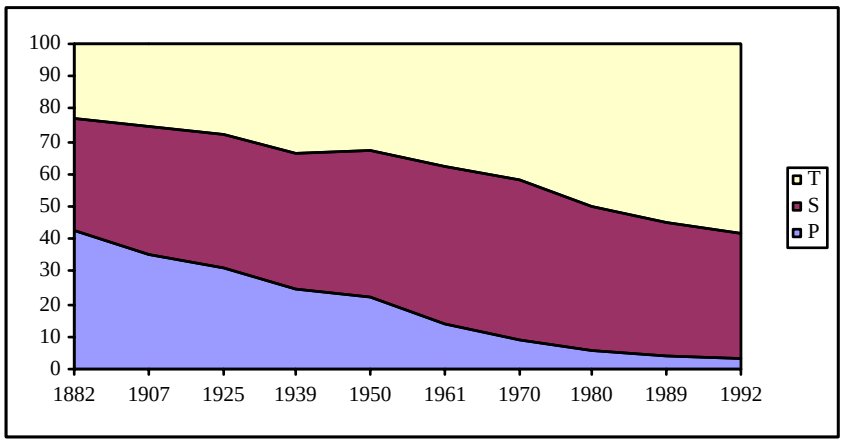
Ano	Sector primário	Sector secundário	Sector terciário
1882	43	34	23
1907	35	40	25
1925	31	41	28
1939	25	41	34
1950	22	45	33
1961	14	48	38
1970	9	49	42
1980	6	44	50
1989	4	41	55
1992	3	39	58

Sector primário: agricultura, pesca; Sector secundário: manufactura e construção
Sector terciário: comércio e todos os outros serviços

Comente a evolução verificada.

Observação: Em termos económicos as sociedades tradicionais empregam grande percentagem da sua força de trabalho no sector primário, enquanto que as sociedades mais industrializadas e mais desenvolvidas têm um maior investimento do capital humano no sector terciário.

Uma representação gráfica possível para mostrar a evolução entre os diferentes sectores da economia, ao longo dos anos é a seguinte:



Tendo em conta a representação gráfica considerada, diga qual o sector que predominava em 1961? E em 1989?

À luz do observação considerada anteriormente, quais dos seguintes países da Comunidade Europeia estão particularmente desenvolvidos (indique os três mais desenvolvidos) e quais os que estão menos desenvolvidos (indique os três menos desenvolvidos)?

País	Sector primário: agricultura		Sector secundário: indústria		Sector terciário: serviços	
	1980	1989/91	1980	1989/91	1980	1989/91
Alemanha(Ocid)	6	4	44	40	50	56
Bélgica	3	3	36	28	61	69
Dinamarca	7	6	32	27	61	67
Espanha	17	11	37	34	49	55
França	9	6	35	29	56	65
Grécia	31	24	29	28	40	48
Holanda	6	5	32	25	62	70
Irlanda	19	15	34	28	47	57
Itália	12	9	41	32	47	59
Luxemburgo	5	3	35	31	60	66
Portugal	26	18	37	34	37	48
Reino Unido	3	2	38	29	59	69

2.1.4. Exemplo 5 (Freedman, 1991) – Idade de indivíduos adultos

A tabela seguinte mostra a distribuição das frequências (em %) relativas do último dígito das idades dos indivíduos adultos. Esta informação foi recolhida relativamente a dois censos diferentes: o Censo de 1880 e o de 1970.

Dígito	*1880*	*1970*
0	16.8	10.6
1	6.7	9.9
2	9.4	10.0
3	8.6	9.6
4	8.8	9.8
5	13.4	10.0
6	9.4	9.9
7	8.5	10.2
8	10.2	10.0
9	8.2	10.1

a) Da consulta da tabela verifica a existência de algumas anomalias?

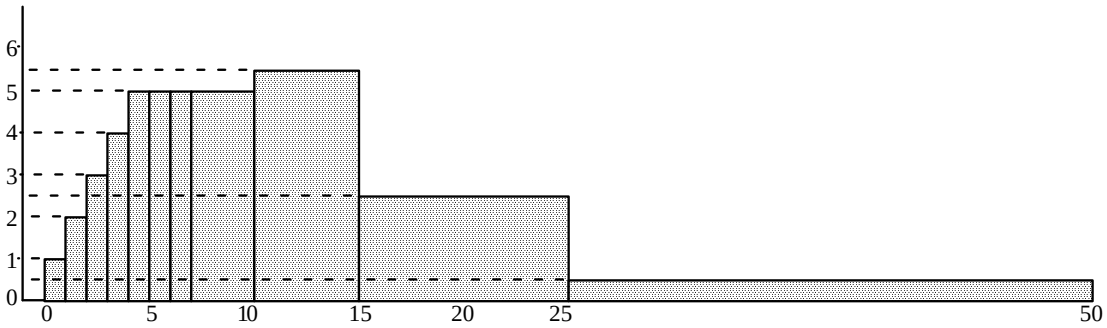
b) Construa diagramas de barras relativamente aos dois censos.

c) Em 1880 havia uma nítida preferência pelos dígitos 0 e 5. Tem alguma explicação para este facto?

d) Em 1970 essa preferência é muito mais fraca. Como explica esse facto?

2.1.5. Exemplo 6 (Freedman, 1991) – Rendimento das famílias americanas

O histograma seguinte representa o rendimento familiar, em milhares de dólares de famílias americanas em 1973.



Cerca de 1% das famílias têm rendimentos entre 0 e 1000 USD. Estime a percentagem de famílias com rendimentos:

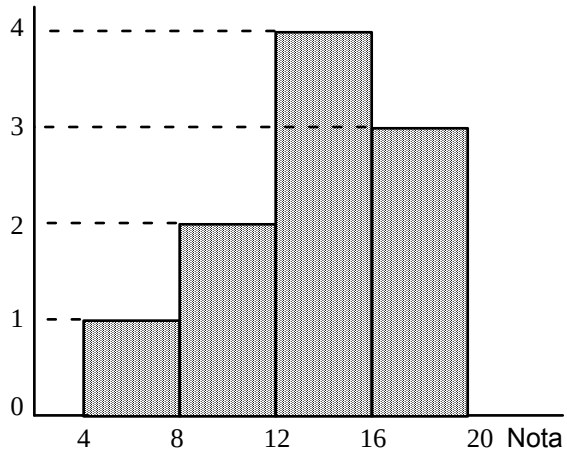
i) a) Entre 1000 USD e 2000 USD
b) Entre 2000 USD e 3000 USD
c) Entre 3000 USD e 4000 USD
d) Entre 4000 USD e 5000 USD
e) Entre 4000 USD e 7000 USD
f) Entre 7000 USD e 10000 USD

ii) a) Haverá mais famílias com rendimentos entre 6000 USD e 7000 USD ou entre 7000USD e 8000 USD? Ou será aproximadamente o mesmo?
b) Haverá mais famílias com rendimentos entre 10000 USD e 11000 USD ou entre 15000USD e 16000 USD? Ou será aproximadamente o mesmo?
c) Haverá mais famílias com rendimentos entre 10000USD e 12000USD ou entre 15000USD e 20000USD?

R: i) a) 2% b) 3% c) 4% d) 5% e) 15% f) 15%
ii) a) O mesmo b) Mais entre 10000 USD e 11000 USD
◊ Mais entre 15000USD e 20000USD

2.1.6. Exemplo 7 – Distribuição das notas a Matemática de uma turma

O histograma seguinte mostra a distribuição das notas finais de Matemática de uma determinada turma.



a) Algum aluno teve nota inferior a 4?

b) Sabe-se que 10% dos alunos da turma tiveram nota entre 4 e 8. Qual a percentagem de alunos com nota entre 8 e 12?

c) Qual a percentagem de alunos com nota superior a 12?

2.1.7. Exemplo 8 (Tannenbaum, 1998) – Salários auferidos no primeiro emprego

Na seguinte tabela de frequências e respectivo histograma estão representados os salários (em dólares) auferidos no primeiro emprego de 3258 formandos na Tasmania State University:

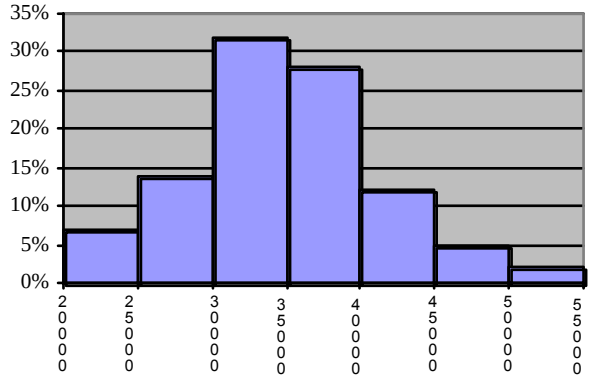
a) Qual a percentagem de alunos com salário inferior a 30000 USD?

b) Qual a percentagem de alunos com salário igual ou superior a 45000 USD?

c) A partir da representação gráfica, diga se há mais alunos com salário entre 30000 e 35000 USD ou entre 35000 e 40000 USD?

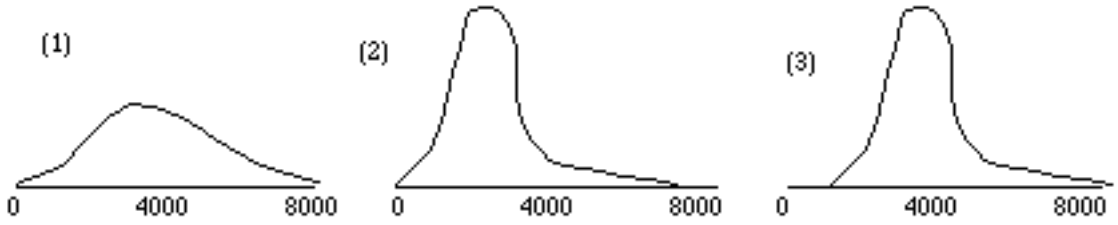
d) Dê um valor aproximado para o salário S tal que 50% dos alunos tenham um salário menor ou igual a S e os restantes alunos tenham um salário maior ou igual a S.

Salário	Freq.abs.	Freq.rel.
[20000, 25000[	228	7%
[25000, 30000[	456	14%
[30000, 35000[	1043	32%
[35000, 40000[	912	28%
[40000, 45000[	391	12%
[45000, 50000[	163	5%
[50000, 55000[	65	2%
Total	3258	100%



2.1.8. Exemplo 9 (Freedman, 1991) – Preços, por hora, de três tipos de trabalho

Recolheram-se os preços, por hora, de 3 tipos de trabalhadores. Os trabalhadores do grupo B ganham cerca de duas vezes mais do que os trabalhadores do grupo A; os trabalhadores do grupo C ganham mais 1500$ por hora do que os do grupo A. Qual das manchas seguintes, de histogramas, pertence a cada um dos grupos?



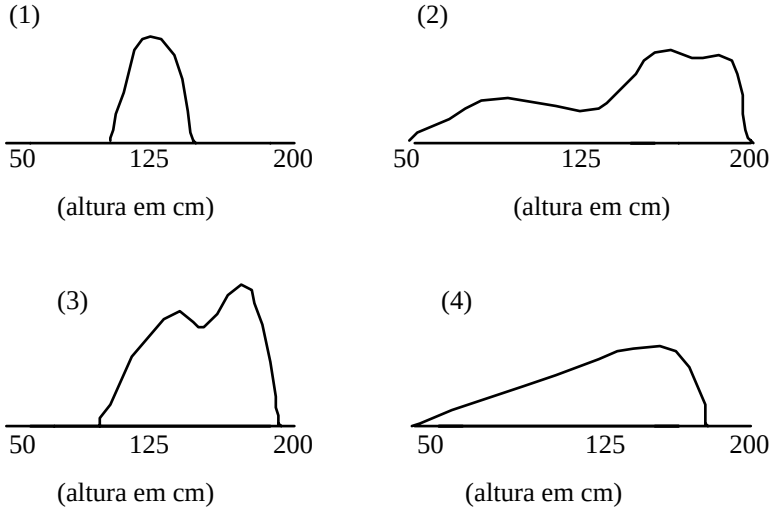
R: (1) - B (2) - A (3) – C

2.1.9. Exemplo 10 (Freedman, 1991) – Alguns exemplos de histogramas

Seguidamente apresentam-se 6 "manchas" de histogramas, 4 dos quais apresentam os resultados do estudo, numa pequena cidade, das 4 características seguintes :

a) Alturas de todos os elementos das famílias, em que os pais tenham idade inferior a 24 anos.
b) Alturas dos casais (marido e mulher).
c) Alturas de todos os indivíduos da cidade.
d) Alturas de todos os automóveis.

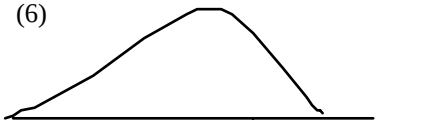
Quais dos histogramas podem representar cada uma das variáveis anteriores? Explique porquê.



(5)

(peso em kg)

(6)



(rendimento em escudos por dia)

R:a) - (2) b) - (3) c) - (4) d) - (1)

2.1.10. Exemplo 11 – O diagrama de caule-e-folhas para comparar dois conjuntos de dados

Considere o seguinte diagrama em caule-e-folhas para comparar os resultados (numa escala de 0 a 100) de duas turmas, no mesmo teste:

Classe 1										Classe 2								
								4	4*									
	4	3	3	3	3	2	2	1	4.									
	8	8	6	5	5	5	5	5	5*									
9	4	4	3	2	2	1	0	0	5.	6								
4	9	8	7	6	5	5	5	5	6*									
			3	3	3	2	2	1	6.	7	8							
				8	8	6	6	5	7*	0	0	0	1					
						4	3	2	7.	5	5	7	8	9	9			
								6	8*	0	0	0	1	1	2	2	2	4
									8.	5	5	5	5	6	6	7	8	
									9*	0	0	1	4					
									9.	6								

Compare os resultados das duas turmas.

2.1.11. Exemplo 12 – Mensagem alarmista (dados hipotéticos)

Numa reportagem de um Telejornal de uma estação de televisão, em princípios de Janeiro, chamava-se a atenção para o aumento da criminalidade na cidade SEMNOME, nomeadamente no que dizia respeito a crimes violentos. Comentava-se que do ano de 1998 para o ano de 1999, a percentagem de crimes violentos tinha aumentado de 17.4%, já que tinha passado de 466 para 547. A reportagem desenvolvia o tema sobre a falta de eficácia da polícia e do Governo no combate ao crime. Considere os seguintes dados relativos à população e ao nº de crimes violentos na referida cidade, nos últimos 6 anos:

Ano	População	Crimes violentos
1994	28650	372
1995	32570	392
1996	36567	405
1997	42456	424
1998	46550	466
1999	55789	547

a) Calcule, para cada ano, a percentagem de crimes violentos, relativamente à dimensão da população.

b) Concorda com o teor da reportagem considerada. Explique porquê.

2.2. Planeamento e aquisição de dados. Questões éticas relacionadas com as experimentações. Exemplos.

Objectivos a atingir:

- ✓ Apresentar as ideias básicas dos processos conducentes à recolha de dados válidos.
- ✓ Fazer sentir a necessidade de aleatoriezar os processos de recolha de dados.

Neste módulo, que consideramos de grande importância, é que se tem a oportunidade de dar a entender o que é a Estatística, como ciência. Em qualquer procedimento estatístico estão, de um modo geral, envolvidas duas fases importantes, nomeadamente a fase que diz respeito à organização dos dados – Análise de dados, e a fase em que se procura retirar conclusões a partir dos dados, dando ainda informação de qual a confiança que devemos atribuir a essas conclusões – Inferência Estatística. Existe no entanto uma fase pioneira, que diz respeito à Produção ou Aquisição de Dados. Como é referido em Tannenbaum et al. (1997), pag 426, "Behind every statistical statement there is a story, and like any story it has a beginning, a middle, an end, and a moral. In this first statistics chapter we begin with the beginning, which in statistics typically means the process of gathering or collecting data. Data are the raw material of which statistical information is made, and in order to get good statistical information one needs good data".
Neste módulo deve-se começar por, face a um determinado problema, identificar a **População** sobre a qual se pretende recolher informação. Depois de identificada devidamente a População é necessário planear cuidadosamente o que é que se pretende medir nos indivíduos que a constituem. De realçar que sobre uma População podemos estar interessados em medir mais do que uma característica populacional ou variável (característica que possa assumir valores ou modalidades diferentes de indivíduo para indivíduo). De um modo geral, não se examina a população toda, mas uma parte a que damos o nome de **Amostra**. De seguida, e de um modo geral, pretendemos retirar conclusões para a População a partir do estudo da Amostra, pelo que a selecção dos indivíduos da População – *amostragem* - sobre os quais vamos efectuar as medições de modo a *produzir* dados – *sondagem* - deve ser feita de modo a obter uma amostra representativa. Deve ser referido que a sondagem visa estudar características da população tal como ela se apresenta.

Devem ser exemplificadas boas e más técnicas de recolha de amostras. De entre as más técnicas realçam-se as *Amostragens por Conveniência* e *as Amostragens por Resposta Voluntária*, técnicas largamente utilizadas, nomeadamente pelos meios de comunicação

social. De entre as boas técnicas realça-se a *Amostragem Aleatória Simples*, a *Amostragem Estratificada* e a *Amostragem Sistemática*.

Incentivar a utilização da máquina de calcular ou de uma folha de cálculo, para gerar números pseudo-aleatórios, para proceder à recolha de amostras aleatórias simples. Do mesmo modo incentivar a utilização da folha de cálculo para a recolha de uma amostra sistemática.

Este assunto da recolha de uma Amostra, com o objectivo de estudar algumas quantidades desconhecidas – *parâmetros* - da População de onde a Amostra foi retirada, através de quantidades calculadas a partir dos dados da Amostra – *estatísticas* - será retomado no ano seguinte, na secção Inferência Estatística.

A recolha de dados através de sondagens não é suficiente quando se pretende estudar o efeito ou resposta de um conjunto de indivíduos a determinado estímulo ou tratamento (termo utilizado em estatística). Somos assim conduzidos a um outro processo de aquisição de dados que é a *experimentação*. Ao contrário de uma sondagem, numa experimentação impõe-se um tratamento a indivíduos com o objectivo de medir a resposta a esse tratamento. Este processo é largamente utilizado em estudos clínicos. Deve ser abordado o problema das questões éticas relacionado com as experimentações. Por exemplo, no estudo de um novo medicamento para a SIDA, que se pensa curar a doença, como devem ser seleccionados os indivíduos objectos do tratamento?

2.2.1. Exemplo 1 – População e amostra

Identifique, no que se segue, População e Amostra:

a) Salários mensais, auferidos pelos empregados de uma empresa;
b) Notas obtidas a Matemática pelos alunos do 10º ano de uma escola secundária;
c) Idades de 45 alunos do 10º ano, de uma escola secundária;
d) Quantidades de vinho obtidas por 10 agricultores da região do Alentejo;
e) Salários mensais auferidos por 250 empregados na indústria têxtil;
f) Notas obtidas a Português, na 1ª chamada nos exames nacionais de 1999;
g) Quantidades de batata consumidas mensalmente em 100 lares portugueses;
h) Um grupo de 20 doentes seleccionados para tomarem um medicamento novo;
i) Número de carros vendidos por cada um dos 5 empregados de um "stand" de venda de automóveis;
j) Número de leitores de 6 jornais diários.

2.2.2. Exemplo 2 – Terá uma revista a aceitação do público?

Uma editora que pretende auscultar a população sobre a aceitação de uma determinada revista que pretende lançar no mercado decide recolher uma amostra a partir do ficheiro disponível na Ordem dos Engenheiros com os nomes dos sócios. Seleccionou aleatoriamente um certo nº de nomes a quem enviou um inquérito a ser respondido com a informação pretendida.

Comente a forma de seleccionar a amostra.

Comentário: O planeamento feito para a recolha da amostra dá origem a *uma amostra enviesada*. Efectivamente, este planeamento tem dois tipos de erros: escolha de uma *amostra por conveniência* (em que é o investigador que escolhe os possíveis elementos que vão pertencer à amostra), ao considerar a lista da Ordem dos Engenheiros, para facilitar a selecção e tem ainda outro tipo de erro, que é o *da resposta voluntária* (em que é o indivíduo que escolhe se responde ou não).

2.2.3. Exemplo 3 (Graça Martins, 1997) – Processo para obter amostras aleatórias simples

Uma escola tem 123 alunos do 10º ano. Pretende-se fazer um estudo sobre os seus projectos quanto ao prosseguimento de estudos superiores. Para isso resolveu fazer-se um inquérito que abranja uma amostra de 25 alunos. Como obter essa amostra?

Processo: Um método elementar consiste em arranjar 123 papéis ou cartões iguais, escrever em cada um o nome de um aluno, meter tudo num saco, misturar bem e extrair 25 papeis.

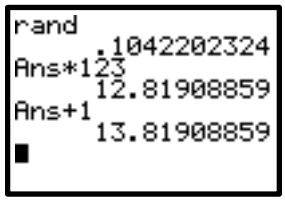
Este método é pouco prático (dá bastante trabalho escrever os 123 nomes) mas funciona bem desde que se tenha o cuidado de misturar cuidadosamente os cartões.

Como quase todas as calculadoras, tanto as científicas simples como as gráficas, possuem uma função geradora de números aleatórios, podemos aproveitar esse facto para um novo método.

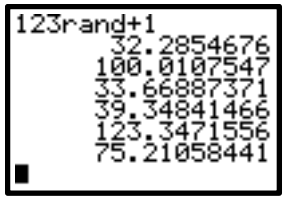
Começamos por numerar os alunos, de 1 a 123.

A função **rand** (ou RND em certas máquinas) gera um número aleatório pertencente ao intervalo [0 ; 1[, intervalo que tem amplitude 1. Podíamos dividir este intervalo em 123 partes iguais e depois ver em qual das partes calhava cada número aleatório que aparecesse. Mas isso não era nada cómodo. Então, o que vamos fazer é arranjar maneira de sortear um número aleatório num intervalo de amplitude 123.

Para isso, poderíamos começar por pedir com **rand** um número aleatório entre 0 e 1. **Multiplicando-o por 123**, passamos a ter um número aleatório pertencente ao intervalo [0 ; 123[. **Somando uma unidade**, o resultado passa a pertencer ao intervalo [1 ; 124[. Se considerarmos só a parte inteira do número obtido, ele vai corresponder exactamente ao número de um dos alunos. No exemplo da figura, seria o aluno nº 13.

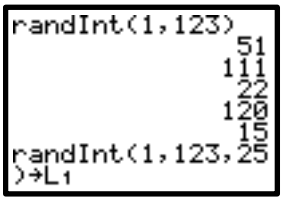


No entanto, podemos fazer isto de forma mais prática escrevendo logo a instrução completa **123** × **rand + 1** , passando a obter um número aleatório pertencente ao intervalo [1 ; 124[cada vez que carregarmos em **ENTER**.

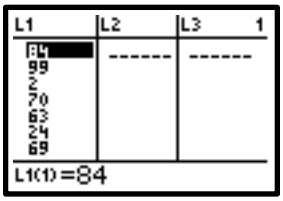


Neste exemplo, os primeiros alunos escolhidos para a amostra são os números 32, 100, 33, 39, 123 e 75. Bastava continuar até obter os 25 elementos, tendo o cuidado de verificar se não surgiam números repetidos.

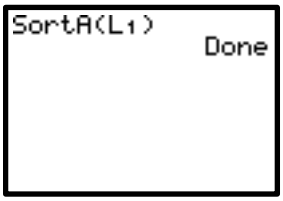
Em certas máquinas, o processo ainda pode ser melhorado do ponto de vista prático com a função **randInt(1,123)** que gera imediatamente um número inteiro aleatório entre 1 e 123 (inclusive).



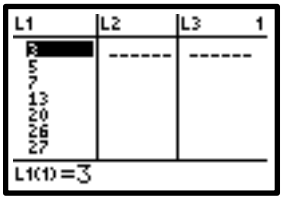
Como queremos 25 números aleatórios, isso pode ser obtido de uma só vez fazendo simplesmente **randInt(1,123,25)** e guardando os números numa lista.



Depois, podemos até ordenar a lista para ser mais fácil ver quais foram os alunos seleccionados.



Contudo, novamente temos de ter o cuidado de verificar se não há números repetidos (e o mais provável é que haja). Se isso acontecer, vai ser preciso sortear mais alguns números.



2.2.4. Exemplo 4 – Recolha de um amostra de professores de Matemática

Suponha que pretende estudar algumas características da População constituída pelos Professores de Matemática que leccionam no Ensino Básico e Secundário, nas escolas públicas, no ano lectivo de 1999-2000.

Diga como poderia seleccionar uma *Amostra* representativa desta *População*, admitindo que dispõe da lista dos professores, fornecida pelo Ministério da Educação.

Resposta: Um processo seria proceder a uma selecção como a exemplificada no exemplo anterior. Outro processo seria o de recolher uma amostra sistemática. Por exemplo, se pretendermos seleccionar uma amostra de 150 professores de uma lista com 6000 professores, considera-se um ficheiro com o nome dos 6000 professores ordenados por ordem alfabética. Considera-se o quociente 6000/150=400 e dos primeiros 400 elementos da lista, selecciona-se um aleatoriamente. A partir deste elemento seleccionamos sistematicamente todos os elementos distanciados de 400 unidades. Assim, se o elemento seleccionado aleatoriamente de entre os primeiros 400, foi o 275, os outros elementos a serem seleccionados são 675, 1075, 1475, etc. Obviamente que o quociente entre a

dimensão da população e a da amostra não é necessariamente inteiro, como anteriormente, mas não há problema pois considera-se a parte inteira desse quociente.

2.2.5. Exemplo 5 – Qual o tempo médio que os alunos da Univ. A gastam diariamente nos transportes?

Pretende-se obter uma estimativa do tempo médio que os alunos de uma Universidade com cerca de 5000 alunos gastam diariamente nos transportes. Sendo a população a estudar relativamente grande decidiu-se seleccionar uma amostra aleatória, de 450 alunos, a quem seria posta a questão. Diga como procederia.

Resposta: Uma vez que os alunos têm um número, faz-se uma selecção utilizando um processo análogo ao considerado no exercício anterior.

2.2.6. Exemplo 6 – Qual a idade média dos alunos da Universidade A?

Considere a situação do exemplo 6, mas admita agora que o que pretende estudar é a idade média dos alunos. Pensa que seria necessário obter uma amostra da mesma dimensão, se pretendêssemos obter resultados dentro da mesma precisão?

Resposta: Bastaria uma amostra de uma dimensão inferior, pois a variabilidade presente nos dados relativamente a esta nova situação é mais pequena.

2.2.7. Exemplo 7 – A dimensão da amostra a recolher tem que ser proporcional à dimensão da população?

Considere ainda a situação do exemplo 6, mas admita agora que a população a estudar tinha 3 vezes mais elementos. Isso significa que para obter resultados com a mesma precisão, necessitava de uma amostra de dimensão 3 vezes superior?

Resposta: Não, porque o tamanho da amostra não tem que ser proporcional à dimensão da população. Como explicava George Gallup, um dos pais da consulta da opinião pública (Tannenbaum, 1998*),: Whether you poll the United States or New York State or Baton Rouge (Louisiana) … you need … the same number of interviews or samples. It´s no mystery really – if a cook has two pots of soup on the stove, one far larger than the other, and thoroughly stirs them both, he doesn´t have to take more spoonfuls from one than the other to sample the taste accurately*".

2.2.8. Exemplo 8 – Relatório Hite (Rossman, 1996)

Em meados dos anos 80 ficou conhecido o relatório Hite, constituído por um estudo das atitudes da mulher relativamente ao seu relacionamento, amor e sexo. Foram distribuídos cerca de 100000 questionários por grupos de mulheres, dos quais foram devolvidos 4500. Destas 4500 mulheres que devolveram os questionários, 96% responderam que davam mais atenção ao marido ou namorado, do que a que recebiam.

Uma sondagem do ABC News/Washington Post feita a 767 mulheres concluiu que 44% lastimavam-se de dar mais atenção do que recebiam.

e) Comente se o relatório Hite pode estar de qualquer modo enviesado e em que direcção. Especificamente acha que o resultado de 96% sobrestima ou subestima a verdade acerca das mulheres americanas?
f) Qual a sondagem que investigou maior número de mulheres?
g) De qual sondagem pensa serem mais representativos da verdade os resultados acerca das mulheres americanas?

2.2.9. Exemplo 9 – Elvis Presley está vivo? (Rossman, 1996)

No vigésimo aniversário da alegada morte de Elvis Presley, uma empresa de Dallas patrocinou uma sondagem a nível nacional. Os ouvintes de mais de 1000 estações de rádio eram convidados a telefonarem para um determinado número (pago) para emitirem a opinião sobre se achavam que Elvis tinha ou não morrido. 56% dos ouvintes disseram que Elvis estava vivo!

Pensa que aquele resultado exprime a opinião da população americana? Identifique alguma fonte de enviesamento na amostra considerada.

2.2.10. Exemplo 10 – Sondagem da SIC sobre a pena de morte

Numa determinada 6ª feira, em que se debateu o aumento de criminalidade a SIC apelou aos telespectadores que respondessem se sim ou não estavam de acordo com a implementação da pena de morte em Portugal, para determinado tipo de crimes. Uma percentagem substancialmente superior a 50% disse que sim. No sábado seguinte o jornal Expresso publicou o resultado de uma sondagem realizada por determinada empresa de sondagens, onde a percentagem de pessoas que eram a favor da pena de morte era consideravelmente pequena, inferior a 10%.

Comente .

2.2.11. Exemplo 11 – Percentagem de mulheres no ensino superior

Segundo fontes do INE, Estatísticas da Educação, o nº de alunos no ensino superior, por sexos, em 1960, 1970, 1981 e 1989 eram

	Homens	%	Mulheres	%	Total
1960	16 839	70.5	7 038	29.5	23 877
1970	25 939	56.4	20 080	43.6	46 019
1981	46 328	55.0	37 845	45.0	84 173
1989	57 879	44.2	73 115	55.8	131 014

Recentemente recolheu-se uma amostra de 500 alunos universitários, tendo-se verificado que 297 eram raparigas. Verifique se a tendência, no que diz respeito à percentagem de alunas no ensino superior, se mantém.

Seguidamente apresentamos alguns casos de estudo, que pela sua relevância, merecem destaque especial.

2.2.12. Caso de estudo 1- A sondagem de 1936 do Literary Disgest (Tannenbaum, 1998)

Nas eleições presidenciais de 1936 nos EUA, defrontaram-se Alfred Landon, o governador republicano do Kansas, e o presidente em exercício Franklin D. Roosevelt. Na altura da eleição a nação não tinha ainda recuperado da Grande Depressão. O Literary Digest, um dos jornais mais respeitados da época, conduziu uma sondagem durante duas semanas antes da eleição. Baseado nesta sondagem o jornal previu que Landon obteria 57% dos votos, contra 43% de Roosevelt. Os resultados da eleição foram 62% para Roosevelt contra 38% para Landon. Como foi possível uma discrepância destas? Na realidade a sondagem levada a cabo pelo Literary Digest foi uma das maiores e mais caras jamais conduzidas, baseada numa amostra de aproximadamente 2.4 milhões de pessoas. Para a mesma eleição a Gallup (Gallup Organization, www.gallup.com) baseada numa amostra muito mais pequena de aproximadamente 50000 pessoas, conseguiu prever a vitória de Roosevelt.

Como foi isto possível?

Comentário: A amostra do Literary Digest foi extraída de uma lista enorme constituída a partir do ficheiro de utentes de telefones, da listagem dos subscritores de jornais e revistas e dos membros das associações profissionais. A partir daí foi criada uma lista de 10 milhões de nomes, tendo sido enviado a cada um, um boletim de voto que deveria ser enviado para o jornal depois de preenchido. Na sua edição de 22 de Agosto de 1936, o Literary Digest apregoava: *Once again, [we are] asking more than then millions voters – one out of four, representing every county in the United States – to settle November's election in October. Next week, the first answers from these ten million will begin the incoming tide of marked ballots, to be triple-checked, verified, five-times cross-classified and totaled. When the last figure has been totted and checked, if past experience is a criterion, the country will know to within a fraction of 1 percent the actual popular vote of forty million (voters).*

A realidade foi bem mais dura! Após a eleição, com a credibilidade completamente desfeita e as vendas em baixo, o Literary Digest foi obrigado a fechar as portas, vítima de um passo em falso estatístico. A primeira coisa que estava errada nesta sondagem foi o processo de selecção para os nomes da lista a quem foi posta a questão, já que esta lista ficou constituída sobretudo por nomes de pessoas das classes média e alta. Em 1936 o telefone ainda era um luxo, assim como o era ser assinante de um jornal ou membro de uma associação profissional, numa altura em que havia 9 milhões de desempregados. Assim a amostra era grandemente *enviesada* e não era de modo nenhum representativa da população. Outro problema a considerar foi o facto de 10 milhões de pessoas terem sido contactadas e só cerca de 2.4 milhões terem respondido. Este problema da não resposta provoca um novo enviesamento, que é muito difícil de corrigir, já que num país livre não se

pode obrigar as pessoas a responder, mesmo pagando, o que não melhoraria a situação, pois introduziria outras fontes de enviesamento.

Moral: É preferível utilizar uma amostra boa, ainda que dimensão pequena, do que uma grande amostra, mas má.

2.2.13. Caso de estudo 2 (Freedman, 1991) – Ensaio clínico da Vacina de Jonas Salk

Em 1916 verificou-se a 1ª epidemia de poliomielite nos Estados Unidos, e durante os 40 anos seguintes esta doença provocou centenas de milhares de vítimas, especialmente crianças. Por volta de 1950, tinham-se descoberto várias vacinas contra esta doença, das quais a que merecia mais confiança era a desenvolvida por Jonas Salk. Efectivamente em experiências laboratoriais mostrou-se eficaz na prevenção e na produção de anticorpos contra a poliomielite. Era, no entanto, necessário conduzir a experimentação fora do laboratório para verificar se a vacina ainda se mantinha eficaz na protecção das crianças.

Em 1954, o Serviço de Saúde Pública decidiu organizar uma experimentação deste tipo. Os indivíduos eram crianças, nas idades escolares mais vulneráveis – níveis 1, 2 e 3. A experimentação seria conduzida em várias escolas de regiões seleccionadas através do país, onde se pensava que o risco de contaminação pela poliomielite era maior. Estavam envolvidas nesta operação 2 milhões de crianças, das quais meio milhão foi vacinada. Deliberadamente não se vacinou 1 milhão de crianças e meio milhão recusaram a vacina. Nesta experiência da vacina de Salk os grupos de tratamento e de controlo têm dimensões diferentes, mas esse facto não traz problemas. Os investigadores comparam as taxas de contaminação pela poliomielite nos dois grupos – nº de casos por cem mil.

Surge então uma questão de ética médica: *Não deveriam todas as crianças ter sido vacinadas*? O problema é que quando se está perante um medicamento novo, mesmo depois de testes laboratoriais extensivos, não é certo que os benefícios compensem os riscos! Tem de se estudar o comportamento do medicamento numa situação real.

Poderíamos pensar que bastaria dar a vacina a um grande nº de crianças, mesmo sem ter um grupo de controlo, pois se por exemplo em 1954 a incidência da poliomielite descesse consideravelmente, relativamente a 1953, então seria a prova da eficácia da vacina de Salk. No entanto, isto não é necessariamente verdade, já que a poliomielite é uma doença epidémica, cuja incidência varia grandemente de ano para ano. Em 1952 houve 60000 casos, enquanto que em 1953 só houve cerca de metade! Sem um grupo de controlo, uma fraca incidência em 1954 poderia significar uma de duas coisas: ou a vacina era eficaz ou nesse ano não houve epidemia.

O único processo de verificar se a vacina era boa, seria deixar algumas crianças sem vacina. Evidentemente que as crianças só seriam vacinadas com autorização dos pais, pelo que

uma possível condução da experiência seria a de formar o grupo tratamento pelas crianças cujos pais consentiram na vacina, enquanto que o outro grupo seria constituído por crianças cujos pais não consentiam. No entanto, sabe-se que é mais fácil obter o consentimento entre as classes socialmente mais favorecidas, do que entre as classes desfavorecidas, o que iria provocar um enviesamento na amostra: embora pareça paradoxal, o nível de incidência da poliomielite é maior no primeiro grupo do que no segundo. O que acontece, é que esta doença está relacionada com a higiene e nas classes mais desfavorecidas, em que as crianças vivem em piores condições higiénicas, estas tendem a contrair casos muito ligeiros de poliomielite, enquanto ainda estão protegidas pelos anticorpos das mães. Por sua vez, a infecção provoca ela própria a criação de anticorpos, que protegem as crianças de casos mais graves da doença.

Assim, para evitar o enviesamento, os grupos de tratamento e de controlo devem ser tão semelhantes quanto possível, para que a diferença nos resultados seja atribuída unicamente ao tratamento e não a outros factores exteriores, cujos efeitos iriam confundir-se com os efeitos do tratamento.

Para a experimentação da vacina de Salk foram propostos vários planeamentos. A National Foundation for Infantile Paralysis (NFIP) propôs vacinar todas as crianças de nível 2, cujos pais consentissem, deixando as crianças de níveis 1 e 3 como grupo de controlo. Este plano foi aceite por muitos distritos escolares. Contudo a poliomielite é uma doença contagiosa, que se propaga por contacto. Assim a incidência pode ter sido maior entre as crianças de nível 2 do que entre as de nível 1 ou 3, provocando um enviesamento contra a vacina. Pode no entanto ter-se verificado o contrário, sendo a incidência mais fraca entre as crianças de nível 2, provocando ainda um enviesamento, mas agora a favor da vacina. Além disso, já que as crianças vacinadas tinham tido o consentimento dos pais, ao contrário das do grupo de controlo, temos novamente um enviesamento contra a vacina, pois o grupo de tratamento inclui demasiadas crianças dos níveis sociais superiores, de acordo com o que dissemos anteriormente.

Muitos distritos escolares atentos a estes problemas existentes no plano NFIP, utilizaram um planeamento diferente. Decidiram que o grupo de controlo tinha de ser escolhido também de entre as crianças, cujos pais tinham dado o consentimento para a vacinação. O problema que se seguia era o de como escolher cada criança para pertencer ao grupo de controlo ou de tratamento. O processo seguido, objectivo e imparcial, consistiu em atribuir cada criança a um dos grupos conforme saísse cara ou coroa, no lançamento de uma moeda equilibrada. Estamos perante uma *experimentação aleatória*.

Uma outra precaução básica, consistiu em usar um *placebo*. Às crianças do grupo de controlo, foi dada uma injecção de água salgada, sem qualquer efeito terapêutico. Durante a

experimentação os indivíduos não sabem se pertencem ao grupo de controlo ou de tratamento, pelo que o resultado da experimentação é unicamente devido ao tratamento e não à "ideia" do tratamento.

Houve ainda outra preocupação, que consistiu no seguinte: os médicos encarregados de verificarem as crianças, não sabiam a que grupo elas pertenciam. O que se passa é que muitas formas da doença são difíceis de diagnosticar, pelo que o diagnóstico poderia ser influenciado pelo facto de se saber que a criança tinha sido vacinada - assim a experimentação é **duplamente aleatória**.

Quais os resultados obtidos?

Na tabela seguinte apresentamos os resultados obtidos para os dois tipos de planeamento considerados, o estudo da NFIP e o outro estudo onde se considera a experimentação com controlo aleatório:

Tabela 1

	Est. aleatorizado		Estudo NFIP		
	Dimensão	*Taxa*		*Dimensão*	*Taxa*
Tratam.	200 000	28	Nível 2 (vacina)	225 000	25
Control	200 000	71	Nível 1 e 3 (contr)	725 000	54
Não cons.	350 000	46	Nível 2(não cons.)	125 000	44

A tabela anterior mostra que o estudo NFIP apresenta um enviesamento contra a vacina. No estudo aleatorizado, a vacina fez descer a taxa da doença de 71 para 28, enquanto que a redução apresentada no estudo da NFIP é bastante inferior. A principal fonte de enviesamento reside no facto de enquanto o grupo de tratamento só incluir crianças cujos pais consentiram na vacina, o grupo de controlo inclui também crianças que não tiveram o consentimento. Assim, o grupo de controlo não é comparável ao grupo de tratamento.

O planeamento aleatório reduz o enviesamento ao mínimo, pelo que deve ser utilizado, sempre que possível.

Eventualmente poderíamos ainda levantar a seguinte questão: será que a vacina é mesmo eficaz? A descida da taxa de poliomielite não será devida ao acaso? A Estatística tem processos – de Inferência Estatística, que permitem concluir que a probabilidade de isso se verificar é extraordinariamente pequena, o que nos levaria a concluir da eficácia da vacina.

Como consequência do estudo anterior procedeu-se a uma vacinação em grande escala, e hoje em dia pode-se dizer que aquela doença está erradicada dos Estados Unidos.

2.2.14. Caso de estudo 3 – Ensaio clínico sobre o *Clofibrate* (Freedman, 1991)

O Coronary Drug Project, foi uma experimentação aleatoriamente controlada, cujo objectivo era o de estudar o comportamento de 5 medicamentos para a prevenção de ataques do coração. Os indivíduos em estudo eram homens de meia idade, com problemas cardíacos. Dos 8341 indivíduos, 5552 foram seleccionados, aleatoriamente, para pertencerem ao grupo tratamento, enquanto que os outros constituíram o grupo de controlo. Os medicamentos e o

placebo foram administrados em cápsulas idênticas. Os doentes foram seguidos durante 5 anos.

Um dos medicamentos em teste foi o *clofibrate*, que reduz os níveis de colesterol no sangue. Infelizmente este tratamento não salvou quaisquer vidas, já que a taxa de morte no grupo em tratamento foi de 20%, durante o período de followup, enquanto que no grupo de controlo foi de 21%. Uma das razões sugeridas para esta falha foi a de que eventualmente muitos dos doentes não teriam seguido o tratamento. Os doentes que tomaram mais de 80% quer do medicamento, quer do placebo foram chamados de "aderentes" ao protocolo. No grupo de tratamento pelo *clofibrate*, a taxa de mortalidade durante o followup foi só de 15% para os aderentes, comparada com 25% para os não aderentes. Este facto mostra que existe evidência para a eficácia do medicamento. Contudo é preciso tomar cuidado! Estamos perante uma comparação observacional, não experimental – embora os dados tenham sido recolhidos enquanto se desenrolava uma experimentação. Efectivamente, os experimentadores não têm poder de decisão sobre quem adere ou não ao protocolo; os próprios indivíduos é que decidem.

	Clofibrate		*Placebo*	
	Nº	Mortes	Nº	Mortes
Aderentes	708	15%	1813	15%
Não aderentes	357	25%	882	28%
Total	1103	20%	2789	21%

Obs: No grupo dos aderentes falta informação sobre 38 indivíduos no grupo do tratamento e 94 no grupo de controlo.

Se repararmos nos dados do grupo de controlo, verificamos que nos aderentes só 15% é que morreram, comparados com os 28% dos não aderentes. Em conclusão:

. O clofibrate não tem qualquer efeito

. O grupo dos aderentes é diferente dos não aderentes.

Provavelmente os aderentes estão mais preocupados com a sua saúde, tomando mais cuidado consigo próprios, pelo que vivem mais tempo.

No Coronary Drug Project, um dos outros medicamentos em teste foi o ácido de nicotina. Suponha que se obtiveram os seguintes resultados:

	Ácido de nicotina		*Placebo*	
	Nº	Mortes	Nº	Mortes
Aderentes	558	15%	1813	15%
Não aderentes	487	27%	882	28%
Total	1096	20%	2789	21%

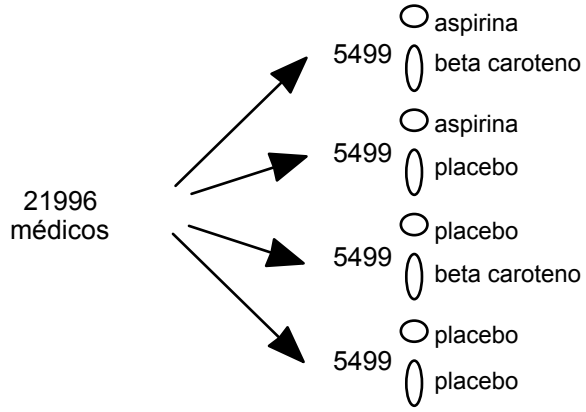
Obs: No grupo dos aderentes falta informação sobre 51 indivíduos no grupo do tratamento e 94 no grupo de controlo.

Alguma coisa parece errada. O quê e porquê?

Verifica-se que a percentagem de aderentes no grupo de tratamento é inferior à do grupo controlo, que são respectivamente 558/1096≈ 51% e 1813/2789 ≈ 65%. Os grupos não são equivalentes. Possivelmente o ácido de nicotina produzirá alguns efeitos secundários, que leva os indivíduos a saírem do tratamento.

2.2.15. Caso de estudo 4 – A aspirina é eficaz na prevenção dos ataques cardíacos?

Ensaio clínico Physicians' Health study (Comap, 2000) - Existe alguma evidência de que tomar regularmente aspirina, em doses baixas, reduz o risco de ataques cardíacos. Suspeita-se também que o mesmo acontece com o *beta caroteno*.O Physicians´Health Study foi um estudo experimental levado a cabo para testar aquelas suspeitas. Envolveu cerca de 22000 médicos do sexo masculino, acima dos 40 anos de idade. Cada um tomou um comprimido todos os dias, durante vários anos. Estavam em estudo 4 tratamentos: aspirina, beta caroteno, ambos e nenhum. No início da experimentação cada médico foi seleccionado aleatoriamente para um dos 4 tratamentos. Neste planeamento está em estudo a ideia do *efeito placebo*. Efectivamente está provado que existe uma tendência para os indivíduos reagirem favoravelmente a qualquer tratamento, mesmo que não tenha qualquer efeito, a não ser psicológico. Por exemplo, se a um dos grupos se desse aspirina e ao outro não se desse nada, qualquer efeito benéfico verificado no grupo que tomou aspirina, pode ser em parte atribuído ao efeito placebo. Assim, é importante que todos os indivíduos envolvidos no estudo tomem comprimidos com o mesmo aspecto e o mesmo sabor, que não lhes permita identificar a qual grupo é que pertencem. Na figura seguinte esquematizamos o planeamento feito:



Por outro lado, os investigadores que conduziram a experiência também não devem saber qual o tratamento a que cada indivíduo foi sujeito, para não influenciar os resultados do exame dos indivíduos em estudo. Este tipo de planeamento diz-se que é *duplamente cego*.

O estudo estatístico dos resultados obtidos (239 ataques cardíacos de entre o grupo que tomou placebo, contra 139 do grupo que tomou aspirina) permitiu concluir que havia evidência para afirmar que a aspirina reduz o risco de ataques cardíacos.

2.3. Aplicação e concretização dos processos anteriormente referidos, na elaboração de alguns pequenos projectos com dados recolhidos na Escola, com construção de tabelas e gráficos simples.

Objectivos a atingir:

- ✓ Fazer sentir a necessidade de organizar os dados, de forma a fazer sobressair a informação neles contida.

Neste módulo pretende-se que os alunos elaborem pequenos estudos em que face a um determinado problema, identifiquem a População objectivo, seleccionem uma amostra representativa, quando não for possível estudar a População toda e façam a redução dos dados obtidos através de uma sondagem. Para organizar os dados devem elaborar tabelas e gráficos, análogos aos já observados no módulo inicial. Nesta fase é importante que o Professor dê a ajuda necessária, quando não for imediata a forma de organizar os dados.

Os projectos efectuados devem estar relacionados com dados recolhidos na Escola ou no meio que rodeia a escola, pois de um modo geral os alunos ficam motivados por estes estudos, já que gostam de conhecer a realidade da sua Escola.

De seguida sugerem-se alguns pequenos projectos, que podem ser realizados por grupos de 4 ou 5 alunos, e que serão apresentados nas aulas, depois de concluídos. Chamamos, no entanto a atenção, que são meras sugestões pois a realidade da Escola poderá sugerir alguns estudos que tenha interesse levar a cabo.

Projecto 1 – Os professores costumam mandar os alunos fazer projectos. Pensa-se que hoje em dia é corrente os alunos terem computador em casa. Será verdade que a maioria dos alunos tem computador em casa? E terão também acesso à Internet?

Projecto 2 – Pretende-se estudar os resultados dos exames nacionais a Matemática e a Português dos alunos que terminaram o secundário no último ano. Verifique nomeadamente a existência de alguma associação entre os resultados a Português e a Matemática dos alunos da escola.

Projecto 3 – Comparar as notas da classificação final interna da disciplina de Matemática com a nota do exame nacional, obtida na mesma disciplina.

Projecto 4 – A distribuição do número de faltas dos Professores faz-se de maneira uniforme para os diferentes dias da semana?

Projecto 5 – Os alunos esperam muito tempo para serem atendidos na fila do bar? Ou na secção de fotocópias? Quais as horas de ponta?

Projecto 6 – Pretende-se planear a construção de um novo campo de jogos na escola. Quais os desportos favoritos dos alunos?

2.4. Classificação de dados. Construção de tabelas de frequência. Representações gráficas adequadas para cada um dos tipos de dados considerados.

Objectivos a atingir:

- ✓ Habilitar na utilização das ferramentas mais adequadas para o tratamento dos diferentes tipos de dados.
- ✓ Ensinar a fazer uma leitura adequada dos gráficos.

Neste módulo procede-se à organização e redução dos dados obtidos através de sondagens ou experimentações. A variável ou variáveis em estudo podem ser de tipo *qualitativo* ou *quantitativo*. Para os dados também se usa a mesma terminologia, conforme resultem da observação de variáveis qualitativas ou quantitativas. Os dados quantitativos ainda podem ser de natureza *discreta* ou *contínua*.

É importante ter presente o tipo de dados objecto de estudo, pois nem sempre se pode aplicar a mesma metodologia estatística a todos os tipos de dados.

Deve ser realçado o facto de as diferentes modalidades que os dados de tipo qualitativo podem assumir, poderem ser representadas por qualquer notação, mesmo numérica. Neste caso, aos números utilizados só se pode eventualmente atribuir um sentido de ordenação e nunca de grandeza associada ao valor do número. Este facto é importante, pois para dados de tipo qualitativo não tem sentido calcular algumas das medidas estatísticas consideradas no módulo seguinte, pois esses dados não se podem adicionar ou multiplicar.

Nesta fase de organização dos dados é essencial construirmos “bons” gráficos, para que tenha sentido a frase vulgarmente utilizada “um gráfico vale mais do que mil palavras”.

Uma das representações gráficas mais simples, com que se pode iniciar este estudo é o *caule-e-folhas*. É uma forma sugestiva de organizar os dados, mas em que se perde pouca informação, pois a maior parte das vezes é possível reconstruir a amostra, só se perdendo a informação da ordem pela qual os dados se apresentavam no conjunto de dados (de maneira geral sem interesse).

Se a representação gráfica *caule-e-folhas* não necessita da construção prévia de uma tabela de frequências, o mesmo não se passa com o *diagrama de barras* – representação mais vulgarmente utilizada para dados qualitativos ou quantitativos discretos, assim como para o *histograma* – representação mais vulgarmente utilizada para dados de tipo contínuo. Assim,

na maior parte das vezes é necessário iniciar a organização de um conjunto de dados construindo uma *tabela de frequências*, onde se apresentam as *frequências absolutas* e as *frequências relativas* e por vezes as *frequências relativas acumuladas*.

A construção de uma tabela de frequências para um conjunto de dados de tipo qualitativo ou quantitativo discreto não apresenta, de um modo geral, dificuldades pois as classes que se consideram são as diferentes modalidades ou diferentes valores que os dados assumem, respectivamente. Para dados de tipo contínuo e por vezes para dados de tipo discreto, é necessário começar por construir classes sob a forma de intervalos, pelo que pode ser dada alguma indicação de quantas classes se devem considerar e de como construir essas classes. No texto de apoio, que acompanha o programa, é dada uma indicação de uma possível regra para o número de classes que de devem considerar, tendo em conta o número de elementos do conjunto de dados a ser tratado, assim como se dão algumas indicações de como devem ser construídas as classes.

Além das representações gráficas referidas anteriormente, será também de considerar o *diagrama circular*, meio vulgarmente utilizado pelos meios de comunicação social para transmitirem a informação contida nos dados.
Deve-se também lembrar que a forma apresentada pelas representações gráficas caule-e-folhas, diagrama de barras ou histograma, reflecte a forma da distribuição da População subjacente aos dados a serem estudados, nomeadamente no que diz respeito à simetria ou assimetria, maior ou menor concentração e existência de valores estranhos (vulgarmente designados de “outliers”).

2.4.1. Exemplo 1 – Classificação de variáveis

Para cada uma das variáveis indicadas a seguir, indique se é de tipo *qualitativo* ou *quantitativo* e neste caso se é de tipo *discreto* ou *contínuo*:

a) Número de calorias de uma sanduíche;
b) Cor dos olhos de uma pessoa;
c) Tempo que uma pessoa leva, de manhã, a ir de casa para o trabalho;
d) Sexo de um indivíduo;
e) Se sim ou não, um estudante vive em casa dos Pais, enquanto estuda;
f) Número de filhos de um casal;
g) Comprimento do salto de um atleta;
h) Estado civil de um indivíduo;
i) Conta de telefone paga mensalmente por uma família;
j) Número de impulsos telefónicos utilizados mensalmente por uma família;
k) Classificação de um automóvel em pequeno, médio e grande;
l) Mês de nascimento de cada estudante de uma dada turma.

2.4.2. Exemplo 2- Estudo dos alunos de uma Escola

A seguinte tabela apresenta as respostas de 38 alunos de uma Escola, a um inquérito, em que se pedia que indicassem: Sexo, Idade, Nº de irmãos, se tinham ou não Cartão de crédito, Altura (cm), Peso (kg) e Desporto preferido:

Sex	Id.	Nº Irm.	Cart.	Alt. cm	Peso kg	Des.	Sex	Id.	Nº Irm.	Cart.	Alt. cm	Peso kg	Des.
M	15	1	S	160	62	Futebol	F	16	0	S	159	45	Ténis
M	14	2	N	162	63	Volei	F	15	4	N	150	46	Basket
F	14	0	N	155	52	Ténis	M	16	2	N	164	58	Vólei
M	16	2	N	164	61	Futebol	F	14	2	S	160	57	Ténis
F	15	3	N	158	50	Andeb.	M	16	3	S	155	46	Nataç.
F	14	1	S	159	51	Ténis	M	15	1	S	157	49	Futebol
F	14	2	S	161	50	Basket	M	15	1	N	163	57	Vólei
F	15	0	N	157	50	Ginást.	F	15	6	N	154	54	Ténis
M	16	1	N	162	61	Futebol	F	16	1	S	156	51	Nataç.
F	16	2	N	160	49	Nataç.	F	14	2	N	158	52	Ginást.
M	15	3	N	163	63	Ténis	M	15	2	S	159	47	Futebol
F	15	4	S	161	49	Basket	F	15	0	N	161	60	Ténis
M	17	0	S	165	65	Ténis	M	14	0	N	162	52	Andeb.
M	15	1	S	162	61	Nataç.	F	16	2	N	159	50	Ginást.
F	16	1	N	155	46	Andeb.	F	15	1	S	160	60	Ginást.
F	15	1	S	154	48	Ginást.	F	14	1	S	156	47	Nataç.
F	14	3	N	156	49	Nataç.	M	15	1	N	162	50	Futebol
M	15	2	S	159	56	Futebol	M	15	2	S	153	51	Ténis
F	14	2	S	157	48	Nataç.	F	16	0	S	157	43	Ténis

a) Classifique as variáveis quanto ao tipo;
b) Construa tabelas de frequências e faça representações gráficas adequadas para os diferentes conjuntos de dados da tabela anterior;
c) Utilizando representações gráficas adequadas, compare os pesos dos rapazes e das raparigas.

Sugestão: Para a alínea b) sugere-se e a construção de diagramas de barras para as variáveis Sexo, Nº irmãos, Cartão e Desporto preferido. Para a variável idade sugere-se a construção de um histograma com as classes [14, 15[, [15, 16[, [16, 17[, [17, 18[. Para as variáveis altura e peso considerar 5 classes de amplitudes iguais. (Por exemplo para a variável altura considerar a amplitude da amostra, isto é a diferença entre o máximo e o mínimo, dividir por 5 o valor obtido e considerar para amplitude classe h um valor aproximado por excesso do resultado da divisão. As classes serão [mínimo da amostra, mínimo da amostra+h[, [mínimo da amostra+h, mínimo da amostra+2h[, [mínimo da amostra+2h, mínimo da amostra+3h[, [mínimo da amostra+3h, mínimo da amostra+4h[, [mínimo da amostra+4h, mínimo da amostra+5h[).

Observação: Quando se pretende construir um histograma de uma amostra de dimensão n, se não houver alguma indicação de quais as classes a constituir, uma regra que costuma dar bons resultados, consiste em considerar para o número de classes k, o menor inteiro tal que.

$2^k \geq n$

2.4.3. Exemplo 3 – Resultados do exame nacional de Português A

Na seguinte tabela apresentam-se os resultados dos 12423 alunos que fizeram exame de Português A - 1ª chamada, tal como nos foi facultada:

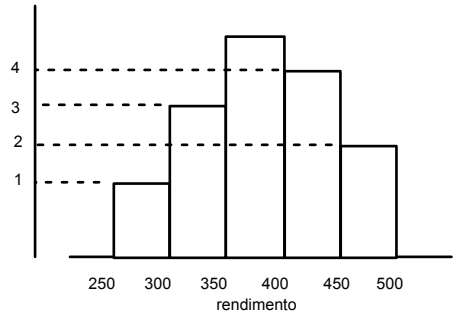
Classe	Nº alunos	Classe	Nº alunos	Classe	Nº alunos
[0,5]	91	]50, 55]	341	]105, 115]	1246
]5, 10]	86	]55, 60]	320	]115, 125]	1021
]10, 15]	54	]60, 65]	443	]125, 135]	825
]15, 20]	45	]65, 70]	399	]135, 145]	630
]20, 25]	81	]70, 75]	606	]145, 155]	437
]25, 30]	91	]75, 80]	452	]155, 165]	283
]30, 35]	154	]80, 85]	663	]165, 175]	177
]35, 40]	163	]85, 90]	339	]175, 185]	62
]40, 45]	224	]90, 95]	1694	]185, 195]	3
]45, 50]	216	]95, 105]	1277		

Faça uma representação gráfica sob a forma de histograma e comente alguns pontos que lhe pareçam de destacar.

Observação: Deve ter em consideração que o histograma é um diagrama de áreas e como tal a área do rectângulo correspondente a cada classe deve ser igual ou proporcional à frequência relativa ou absoluta da classe.

2.4.4. Exemplo 4 – Rendimento familiar dos habitantes numa zona de Lisboa (Exemplo Hipotético)

Tendo sido feito um estudo sobre o rendimento familiar dos residentes em determinada zona da cidade, recentemente construída e habitada fundamentalmente por casais jovens, verificou-se que esse rendimento (em milhares de escudos) se distribuía da seguinte forma:

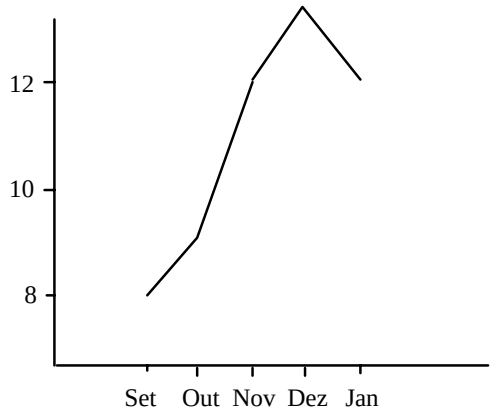


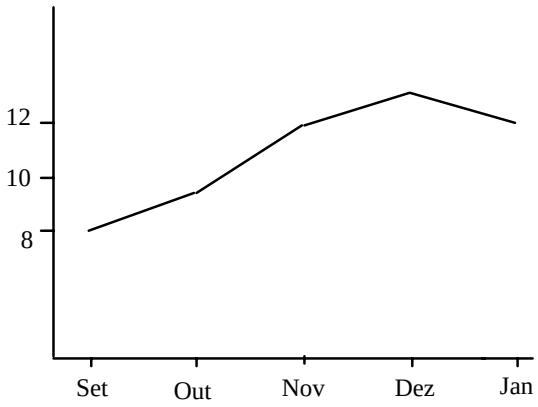
Admitindo que 10% das famílias têm rendimento até 300 contos:

a) Qual a percentagem de famílias com rendimento entre 350 e 400 contos?
b) Qual a percentagem de famílias com rendimento superior a 420 contos?
c) Qual o valor para o percentil 20 (isto é, qual o rendimento máximo auferido pelas 20% das famílias de menores rendimentos)?

2.4.5. Exemplo 5 – Número de acidentes na IP5

(Exemplo Hipotético) - Suponha que o nº de acidentes no IP5 foi, no período de Setembro de 1997 a Janeiro de 1998, o seguinte: 8, 9, 12, 13 e 12. Dois jornais apresentaram as seguintes representações gráficas para transmitirem a informação anterior:



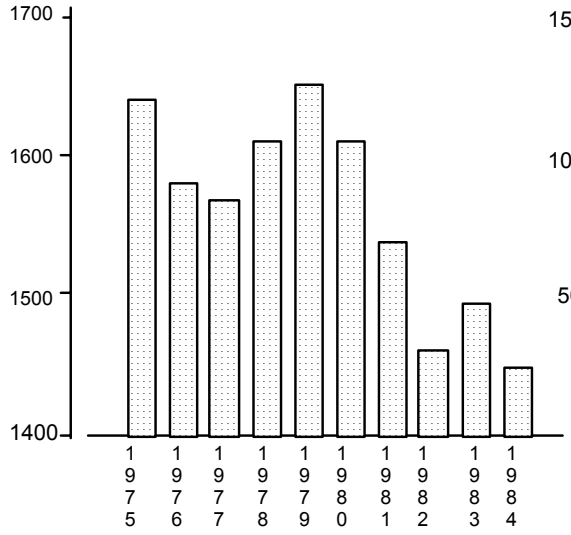


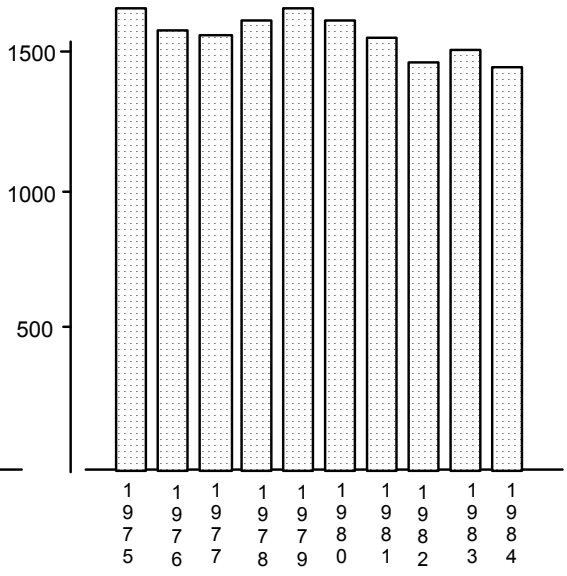
Comente as representações gráficas anteriores.

2.4.6. Exemplo 6 – Diminuição do número de vendas de livros em Portugal

(Exemplo Hipotético) - Os dois diagramas de barras seguintes pretendem traduzir a mesma informação, pois dizem ambos respeito ao número de vendas de livros em Portugal, entre 1975 e 1984:

Qual ddos diagramas traduz mais correctamente a informação?





2.4.7. Exemplo 7 – A condução torna-se mais segura diminuindo a velocidade?

(Moore, 1977) - Em 1970, nos Estados Unidos o Governo decretou uma diminuição no limite de velocidade na estrada para 55 milhas, assim como outras medidas de segurança. Esta decisão fez com que o número de mortes por acidentes diminuísse de 52600, em 1970, para 51091 em 1980. Perante tão pequena diminuição (3%) no número de acidentes, seremos levados a concluir que a condução não se tornou tão segura, quanto se esperava? Precisaria de mais alguma informação para tirar conclusões?

Comentário: O que aconteceu é que o número de veículos registados cresceu de 108 milhões em 1970, para 156 milhões em 1980. Assim o número de mortes não teve em consideração o aumento do número de condutores. Uma medida correcta para indicar a taxa de mortes poderá ser dada pela proporção de acidentes relativamente ao números de carros, ou melhor, pelo número de acidentes relativamente ao número de milhas percorridas. Tendo em conta esta informação, quando se calculam estas proporções verifica-se que a taxa de mortalidade desceu de 4.7 mortes por 100 milhões de milhas em 1970 para 3.3 em 1980, o que significa uma queda de cerca de 30%.

2.4.8. Exemplo 8 – Comparação da Coproporfirina nas mulheres grávidas

A concentração em CPU (coproporfirina urinária) em 35 mulheres grávidas internadas é medida em período diurno e em período nocturno, com intuitos comparativos.

Mulher (i)	diurno	nocturno	Mulher (i)	diurno	nocturno
1	40.8	50.4	19	38.8	61.3
2	42.7	48.2	20	43.3	55.3
3	36.8	59.4	21	34.0	58.2
4	43.5	43.3	22	39.4	42.9
5	35.4	59.6	23	43.4	55.2
6	39.1	60.5	24	38.7	51.6
7	41.8	51.2	25	37.2	41.0
8	41.6	61.6	26	36.7	32.8
9	36.8	44.0	27	37.3	55.0
10	37.0	59.6	28	35.1	45.8
11	43.4	51.8	29	36.0	60.8
12	41.8	67.4	30	41.9	67.8
13	42.5	51.7	31	38.8	50.6
14	35.8	51.0	32	37.4	52.3
15	35.9	52.9	33	38.6	52.8
16	40.9	67.1	34	39.1	45.4
17	41.5	61.7	35	38.9	47.0
18	41.1	45.9			

Considerando separadamente os dados correspondentes ao período diurno e ao período nocturno, represente-os graficamente. Compare as representações gráficas assim obtidas e comente as conclusões a que chegar.

Sugestão: Sugere-se a construção de caules-e-folhas paralelos para comparar os dois conjuntos de dados. Uma representação possível é a que se apresenta a seguir:

Nocturno										Diurno																				
								2	3	4																				
									3	5	5	5	5	6	6	6	6	7	7	7	7	8	8	8	8	8	9	9	9
						4	3	2	1	4	0	0	1	1	1	1	1	1	2	2	3	3	3	3						
					8	7	5	5	5	4																				
2	2	2	1	1	1	1	1	0	0	5																				
			9	9	9	8	5	5	5	5																				
					1	1	1	0	0	6																				
							7	7	7	6																				

Como se verifica na representação gráfica anterior, os valores correspondentes à concentração de CPU, é muito maior no período nocturno do que no período diurno. Por outro lado a variabilidade apresentada no período nocturno também é superior à do período diurno, em que os dados são muito concentrados.

Observação: Na construção do caule-e-folhas anterior truncamos o dígito das decimas.

2.5. Cálculo de estatísticas. Vantagens, desvantagens e limitações das medidas consideradas.

Objectivos a atingir:
Apresentar umas medidas, que tal como as representações gráficas, permitem reduzir a informação contida nos dados.
Chamar a atenção para as vantagens e para as situações em que não se devem calcular.
Além das representações gráficas também se utilizam medidas calculadas a partir dos dados – *estatísticas*. Destas medidas destacam-se as *medidas de localização*, nomeadamente as que localizam o centro da amostra, de que destacamos a *média* e a *mediana*, e *medidas de dispersão*, que medem a variabilidade apresentada pelos dados, de que destacamos o *desvio padrão* e a *amplitude inter-quartil*. Outras medidas de localização a considerar são os quantis, nomeadamente os *quartis* e os *percentis*.
Deve-se observar que ao reduzir a informação contida nos dados sob a forma de alguns números, se está a proceder a uma redução drástica desses dados, pelo que as estatísticas consideradas devem ser convenientemente escolhidas de modo a representarem o melhor possível os dados que pretendem sumariar.
Pelo que ficou referido no parágrafo anterior é importante referir para cada uma das medidas consideradas não só o processo de as calcular, mas também as suas limitações. Um exemplo a realçar é o facto de ter pouco interesse falar em centro de uma distribuição de dados para dados fortemente enviesados e muito menos utilizar a média como medida de localização deste tipo de dados.
Chamar a atenção, através de exemplos, para o facto de a média não ser uma medida *resistente*, por ser muito influenciada pela existência na amostra de valores muito pequenos ou muito grandes (outliers), mesmo que em pequena quantidade. Realçar a importância da mediana – medida resistente, como alternativa à média, para as situações em que esta não deve ser utilizada.
Uma vez exploradas as medidas de localização mediana e quartis, introduzir uma nova representação gráfica – *diagrama de extremos e quartis*, e realçar as suas vantagens, nomeadamente no que diz respeito à simplicidade de construção e à informação que traduz no que diz respeito à simetria e à dispersão dos dados, não só na parte central, mas também nas caudas da distribuição. Realçar ainda a importância desta representação gráfica quando se pretendem comparar vários conjuntos de dados.
Realçar o facto, nomeadamente através de exemplos, de que um conjunto de dados não fica bem caracterizado unicamente através das medidas de localização, sendo necessário utilizar também as medidas de variabilidade ou dispersão. Destas destacar o *desvio-padrão*, com limitações idênticas às da média, no que diz respeito a dados enviesados ou com outliers. Apresentar como alternativa a *amplitude inter-quartil*.

Chamar a atenção para a importância de uma representação gráfica como início de estudo de um conjunto de dados, pois se os dados são aproximadamente simétricos podemos considerar o par de estatísticas média e desvio padrão para os caracterizar, mas se existe um enviesamento considerável então deve ser considerado um conjunto de 5 números, normalmente conhecido por *resumo dos cinco números*, constituído pela mediana, 1º e 3º quartis e extremos, e utilizados na construção do diagrama de extremos e quartis.
Chamar a atenção para uma propriedade, conhecida pela regra dos 68 – 95 –100%, verificada pelos dados que se distribuem de forma aproximadamente "normal", ou seja quando o histograma apresenta uma forma característica com uma classe média predominante e as outras classes distribuindo-se à volta desta de forma aproximadamente simétrica e com as frequências a decrescer à medida que se afastam da classe média. Esta regra será clarificada quando for estudado o Modelo Normal.
Ainda para dados cuja distribuição é aproximadamente "normal" é costume subtrair aos dados a média e dividir pelo desvio padrão, obtendo-se os valores *estandardizados*. Este processo é aconselhado quando se pretendem comparar valores pertencentes a amostras diferentes.
Nesta secção em que se destaca a pouca utilidade do par (média, desvio-padrão) (embora seja o mais divulgado e mais conhecido), para caracterizar distribuições de dados fortemente enviesadas, pode-se falar de *transformações* de dados que permitem reduzir o enviesamento e conduzir a distribuições aproximadamente simétricas.

2.5.1. Exemplo 1 – Atenção com o cálculo das estatísticas

Perguntou-se a cada um dos 80 estudantes de um determinado curso, qual o seu grau de satisfação relativamente ao curso que frequenta. Obtiveram-se os seguintes resultados:

NS	MB	B	S	NS	NS	SP	SP
NS	B	NS	NS	SP	B	B	MB
SP	NS	NS	MB	SP	B	NS	B
SP	S	SP	SP	NS	NS	SP	S
MB	S	B	MB	NS	S	S	S
SP	S	B	NS	S	S	SP	B
B	B	MB	NS	B	S	NS	NS
B	S	MB	S	MB	NS	MB	SP
S	S	NS	B	MB	NS	MB	NS
B	MB	SP	MB	S	SP	SP	MB

NS-"Não Satisfaz"; SP-"Satisfaz Pouco; S-"Satisfaz"; B- "Bom"; MB- "Muito Bom".

Faça uma representação gráfica adequada para os dados e indique as características amostrais que achar conveniente. Substitua as categorias consideradas anteriormente, respectivamente por 1, 2, 3, 4 e 5. Calcule agora as características amostrais que achar convenientes. De que tipo é a variável que está a estudar?

2.5.2. Exemplo 2 – Cálculo de estatísticas.

O cálculo de certas estatísticas pode sugerir a forma da distribuição dos dados?

Considere a tabela do exemplo 2 da secção anterior.

a) Considere a variável sexo e codifique o F com um 0 e o M com um 1. Obtém um conjunto de 0's e 1's. Calcule a média deste conjunto de dados. Agora, proceda a uma nova codificação atribuindo ao F o valor 1 e ao M o valor 2. Obtém um conjunto de 1's e 2's. Calcule novamente a média deste conjunto de dados. Pode dizer que os valores obtidos para a média, representam a média da variável Sexo? Explique.
b) Calcule a média das idades dos rapazes e das raparigas. A partir dos valores obtidos poderá obter a idade dos 38 alunos? Explique como se faz.
c) Calcule o valor da mediana das idades dos alunos. Compare com o valor obtido para a média das idades;
d) Considere a representação gráfica obtida para a característica peso, no exemplo 2 da secção anterior. Tendo em conta a forma do histograma, espera obter para a média um valor superior ou inferior à mediana? Calcule a média e a mediana dos pesos dos alunos e confirme a sua suposição;
e) Calcule o desvio padrão da variável Peso, para os rapazes e para as raparigas. São os rapazes ou as raparigas que apresentam maior variabilidade relativamente a esta característica?

f) Se a distribuição dos dados da variável Peso tivesse um comportamento aproximadamente normal, quantos dados é que esperava obter no intervalo [média dos pesos – desvio padrão dos pesos, média dos pesos + desvio padrão dos pesos]?

g) Qual a moda da variável Desporto preferido? Que outras características amostrais poderá calcular?

2.5.3. Exemplo 3 – A média e a mediana, respectivamente como uma medida não resistente e uma medida resistente. Atenção ao cálculo da mediana.

Considere os seguintes dados que representam o número de mortes de algumas erupções vulcânicas que ficaram célebres (Fonte: World Almanac, 1993):

Data	Nome vulcão	Nº mortes	Data	Nome vulcão	Nº mortes
79 a.c.	Mt. Vesuvius, Italy	16000	1902	Santa Maria, Guatemala	1000
1169	Mt. Etna, Sicily	15000	1902	Mt. Pelée, Martinique	30000
1631	Mt. Vesuvius, Italy	4000	1911	Mt. Taal, Philippines	1400
1669	Mt. Etna, Sicily	20000	1919	Mt. Kelud, Java	5000
1772	Mt. Papandayan, Java	3000	1951	Mt. Lamington, New Guinea	3000
1792	Mt. Unzen-Dake,Japan	10400	1966	Mt. Kelud, Java	1000
1815	Tamboro, Java	12000	1980	Mt. St. Helens, U.S.	60
1883	Krakatau, Indonesia	35000	1985	Nevado del Ruiz, Colombia	22940

a) Calcule a média e mediana do número de mortes. O que pode concluir quanto à simetria da distribuição dos dados?

b) Suponha que ao digitar os valores anteriores o valor que diz respeito à erupção vulcânica de 1883 apareceu 335000, em vez de 35000. Calcule novamente a média e a mediana;

c) Admita agora que o engano se deu ao digitar o 60, que apareceu substituído por 600. Calcule novamente a média e a mediana;

d) Apresente os valores obtidos nas alíneas anteriores no seguinte quadro e comente-o:

	Dados originais	Dados com o valor 335000	Dados com o valor 600
Média			
Mediana			

e) Suponha que um prof. pediu aos seus alunos que calculassem a mediana dos dados respeitantes ao número de mortes, e que alguns apresentaram o valor 18000. O que é que poderá ter acontecido?

2.5.4. Exemplo 4 – A média não é suficiente para caracterizar um conjunto de dados.

Suponha que um prof. fez o mesmo teste a duas turmas tendo seleccionado aleatoriamente 29 e 23 alunos respectivamente da turma 1 e turma 2. Os resultados obtidos são apresentados na tabela seguinte:

Classe	Turma 1	Turma 2
[4, 6[	2	0
[6, 8[	3	3
[8,10[	5	5
[10, 12[	7	6
[12,14[	6	5
[14,16[	4	4
[16,18[	2	0

a) Calcule valores aproximados (com uma casa decimal) para a média das duas turmas e verifique que os valores obtidos são iguais;

b) Os resultados obtidos na alínea anterior permitem-lhe afirmar que as turmas tiveram um comportamento semelhante no teste? Explique porquê.

2.5.5. Exemplo 5 – Uma situação paradoxal causada pela média

a) Considere a tabela 1 onde são representados a esperança de vida, em média, de 40 doentes com cancro do pulmão em 3 fases distintas da doença que vai de 1 (menos grave) até 3 (mais grave):

Tabela 1

Fase	Nºdoentes	Nºanos so - brevivência	média
1	10	8	7
	10	6	
2	10	4	3.5
	10	3	
3	10	2	1.5
	10	1	

Tabela 2

Fase	Nºdoentes	Nºanos so - brevivência	média
1	10	8	?
2	10	6	?
	10	4	
3	10	3	?
	10	2	
	10	1	

Tendo-se descoberto um meio de diagnóstico mais avançado, procedeu-se a uma reclassificação dos doentes, tendo-se obtido a tabela 2. Calcule para cada fase da doença, a média da esperança de vida dos doentes. O que conclui? Não acha a situação paradoxal?

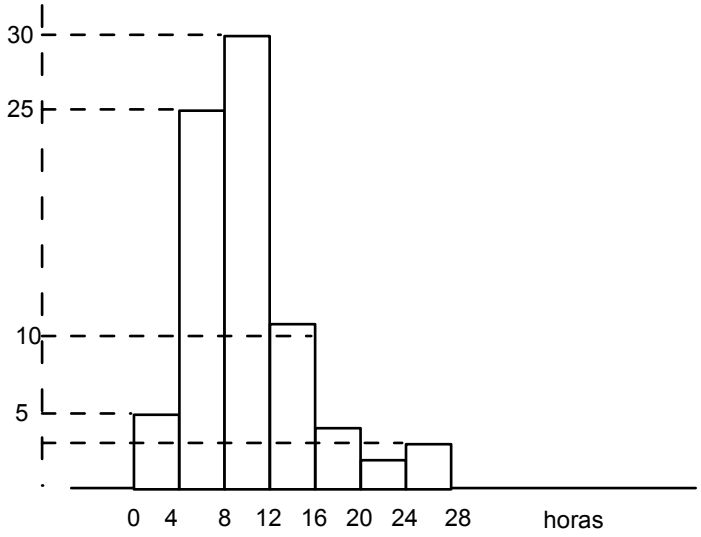
Comentário: Será razoável concluir que um melhor meio de diagnóstico e não um melhor tratamento aumente o tempo médio de sobrevivência?

Calcule a média de sobrevivência para a amostra completa nas duas situações.

b) Will Rogers, um comediante e comentador social dos anos 20 e 30, fez o seguinte comentário, a propósito da emigração em massa que se verificou de Oklahoma para a Califórnia, de pessoas à procura de emprego: “When the Okies left Oklahoma and moved to California, they raised the average intelligence of both states”. Comente a frase anterior.

2.5.6. Exemplo 6 – Cálculo de estatísticas para dados agrupados. Comportamento da média e do desvio padrão para transformações lineares dos dados.

Pediu-se a um certo nº de pessoas que indicassem o tempo médio (em horas) que, por semana, passavam a ver televisão. Com os resultados obtidos construiu-se o seguinte histograma:



a) Admitindo que a percentagem de elementos da amostra que pertencem à classe [0,4[é 6.25%, calcule as percentagens pertencentes às outras classes.
b) Calcule valores aproximados para as seguintes características amostrais: média, desvio padrão e 1º quartil.
c) Se adicionasse 5 pontos a cada um dos elementos da amostra que deu origem ao histograma, como se comporta a média da amostra obtida? E o desvio padrão? Esboce o histograma da nova amostra.
d) Se multiplicasse cada elemento da amostra por 5, qual o novo valor para a média e o desvio padrão?

2.5.7. Exemplo 7 – Comparação de duas amostras

Numa determinada fábrica é necessário que cada operário seja submetido a um período de aprendizagem de um mês, de modo a atingir a eficiência máxima na realização de certa tarefa. Foi realizada uma experiência com o objectivo de comparar o método de aprendizagem habitual, com um novo método. Assim, formaram-se dois grupos de 9 operários cada um, que depois de seguirem os cursos, durante o tempo estabelecido, foram solicitados a realizar a tarefa, para a qual tinham sido treinados. Registaram-se os tempos (em minutos), que se apresentam a seguir:

Método	Tempos
Habitual	32 37 35 28 41 44 35 31 34

Novo 35 31 29 25 34 40 27 32 31

Os dados sugerem uma diferença entre os dois métodos?

2.5.8. Exemplo 8 – Comparação de 4 processos de fabrico (Rossman, 1996)

Num processo de fabrico de aros de aço, pretende-se que o diâmetro dos aros seja de 12 cm, mas o que acontece é que há sempre uma pequena flutuação à volta deste valor. Consideram-se não defeituosos os aros cujo diâmetro esteja no intervalo [12 ± 0.2]. Suponha que se recolheram 50 aros para inspecção produzidos por 4 máquinas, tendo-se obtido os seguintes resultados:

Máq. A	Máq. B	Máq. C	Máq. D	Máq. A	Máq. B	Máq. C	Máq. D
11.5378	12.3179	11.9372	12.1122	11.4882	12.0150	11.7551	12.0107
11.4454	11.6369	11.9516	11.8743	11.5391	12.4587	12.0131	12.1386
11.4973	11.9653	11.6774	11.9922	11.4908	12.0230	11.7715	12.0782
11.4806	11.9294	11.6848	11.9810	11.5439	12.3769	11.9613	12.1253
11.5251	12.0735	11.7001	12.0227	11.4235	11.6453	11.5309	11.8827
11.4221	11.7086	11.5958	11.9032	11.4576	11.8803	11.6867	11.9627
11.5107	12.0805	11.7025	12.0116	11.4208	11.6528	11.5561	11.8888
11.5657	12.5956	12.1176	12.1953	11.4408	11.7646	11.6197	11.9266
11.4678	12.0636	11.8339	12.0202	11.5007	12.1889	11.8806	12.0628
11.4818	11.9256	11.6309	11.9640	11.4828	11.7945	11.5535	11.9390
11.4988	12.2369	11.9200	12.0744	11.4839	12.0097	11.7472	12.0045
11.4165	11.4654	11.3773	11.8257	11.5457	12.1448	11.7488	12.0559
11.5538	12.3717	11.9023	12.1138	11.5636	12.4386	11.9927	12.1520
11.5097	12.3041	11.9396	12.0898	11.4733	11.8538	11.6834	11.9745
11.4662	11.9887	11.7716	11.9980	11.4608	11.5233	11.3408	11.8500
11.4398	11.6219	11.4825	11.8800	11.4485	11.7976	11.6320	11.9369
11.4478	11.7303	11.6008	11.9260	11.4585	11.6921	11.4837	11.8958
11.5635	12.3644	11.8640	12.1089	11.4806	12.1321	11.8958	12.0513
11.5397	12.1163	11.7297	12.0446	11.5222	12.0173	11.6354	11.9988
11.4753	11.8964	11.5965	11.9472	11.4636	11.7704	11.5487	11.9216
11.5313	12.1731	11.7979	12.0604	11.5088	12.1286	11.7875	12.0384
11.5102	11.9675	11.6818	12.0044	11.4551	11.9278	11.7457	11.9803
11.5348	12.1768	11.7582	12.0501	11.5165	12.3054	11.9466	12.0977
11.4574	11.5570	11.3846	11.8618	11.5358	12.1318	11.7388	12.0444
11.5077	11.9915	11.6969	12.0073	11.5402	12.4835	12.0710	12.1588

Descreva cada um dos processos de fabrico, tendo em conta a representação gráfica obtida. Tendo em conta as características das distribuições obtidas, nomeadamente no que diz respeito à localização do centro e variabilidade, responda às seguintes questões:

a) Qual a melhor máquina?
b) Qual a máquina mais estável, isto é, que produz aros com menor variabilidade no diâmetro?
c) Qual o processo menos estável?
d) Qual a máquina que produz aros que de um modo geral têm o diâmetro mais afastado do objectivo?

2.6. Introdução gráfica à análise de dados bivariados quantitativos

Objectivos a atingir:

- ✓ Apresentar um modo eficaz de visualizar a associação entre duas variáveis.
- ✓ Saber interpretar o tipo e a força com que duas variáveis se associam.

Pode acontecer que sobre um indivíduo da população a estudar se recolha informação sobre duas características ou variáveis quantitativas, obtendo assim um conjunto de dados sobre a forma de pares de dados. Normalmente o que se pretende neste caso é estudar a relação entre as duas variáveis, que se supõe estarem relacionadas. O processo adequado para descrever esta relação é começar pela representação gráfica conhecida por *diagrama de pontos* ou *diagrama de dispersão*. O que se pretende retirar de uma representação deste tipo é a forma, direcção e grau de associação entre as variáveis.

Devem ser exemplificadas as diferentes situações que podem surgir, reflectindo os diferentes tipos e graus de associação que se pode verificar entre as variáveis. Nomeadamente quando se fala em associação, pode-se referir que esta pode ser *positiva* ou *negativa* e deve-se precisar o que se entende por associação. Assim, deve-se referir que uma associação positiva significa que, em média, quando uma variável aumenta a outra também aumenta, enquanto que uma associação negativa significa que, em média, quando uma variável aumenta, a outra diminui.

Se se concluir que tem sentido falar numa associação entre as variáveis, traduzida pela nuvem de pontos num diagrama de dispersão, com a forma de uma oval, mais ou menos alongada, então passa-se a uma fase posterior, da construção de um modelo que permita conhecer como se reflectem numa das variáveis as modificações processadas na outra, o que conduzirá aos *modelos de regressão*, a estudar a seguir.

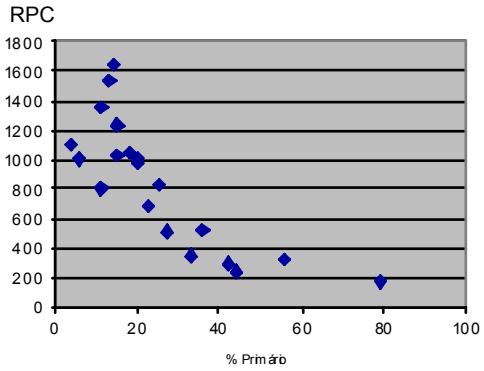
2.6.1. Exemplo 1 – Rendimento per capita e percentagem de força laboral. Aplicação na agricultura, na indústria e nos serviços para 20 países da OCDE, em 1960 (Fonte: lib.stat.cmu.edu)

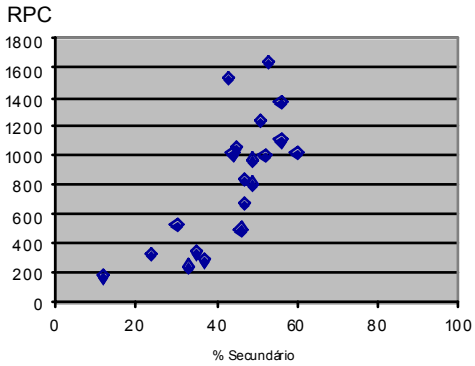
Na seguinte tabela são apresentados o rendimento *per capita* (RPC) de 20 países OCDE, assim como a percentagem da sua força laboral aplicada no sector primário (agricultura), secundário (indústria) e terciário (serviços), em 1960:

País	RPC	Primário	Secundário	Terciário
Canadá	1536	13	43	45
Suécia	1644	14	53	33
Suiça	1361	11	56	33
Luxemburgo	1242	15	51	34
Reini Unido	1105	4	56	40
Dinamarca	1049	18	45	37
Alemanha	1035	15	60	25
França	1013	20	44	36

Bélgica	1005	6	52	42
Noruega	977	20	49	32
Islândia	839	25	47	29
Holanda	810	11	49	40
Áustria	681	23	47	30
Irlanda	529	36	30	34
Itália	504	27	46	28
Japão	344	33	35	32
Grécia	324	56	24	20
Espanha	290	42	37	21
Portugal	238	44	33	23
Turquia	177	79	12	9

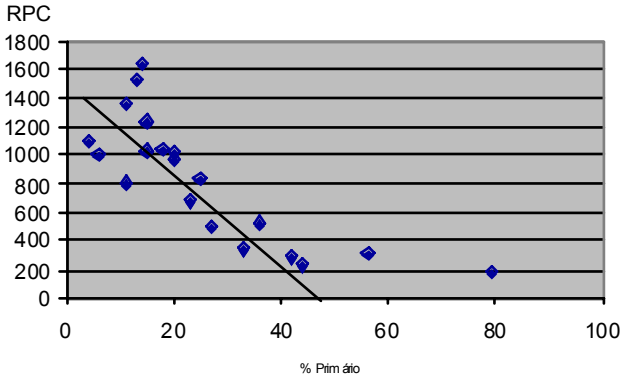
Representando em dois gráficos os pontos de coordenadas (RPC, % Primário) e (RPC, % Secundário) para os diferentes países considerados, obtemos os seguintes diagramas de pontos, que passamos a analisar:





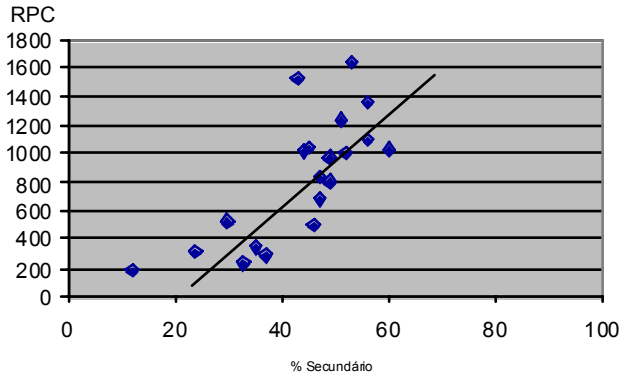
O aspecto apresentados pelos dois diagramas de pontos ou *diagramas de dispersão* (scatter diagrams) é completamente diferente. Assim, enquanto que as variáveis rendimento per capita (RPC) e % da força laboral se associam *negativamente*, isto é, quanto maior é a % de força laboral empregue no sector primário, menor é o rendimento per capita, no que diz respeito às variáveis rendimento per capita e % de força laboral empregue no sector secundário, verifica-se uma associação *positiva*, isto é, quanto maior for a % de força laboral empregue no sector secundário, maior é o rendimento per capita.

Verificamos ainda que, relativamente às variáveis % de força laboral e rendimento per capita empregue no sector primário, o padrão da nuvem de pontos pode ser aproximadamente modelado por uma recta, havendo no entanto 4 países que contrariam esta aproximação por um modelo linear: o Canadá, a Suécia, a Grécia e a Turquia.



O declive da recta é negativo, o que traduz a associação negativa existente entre as variáveis representadas.

No que diz respeito às variáveis % de força laboral empregue no sector secundário e rendimento per capita, a nuvem de pontos segue um padrão aproximadamente linear, como se exemplifica na figura a seguir, havendo no entanto 3 pontos, correspondendo aos países



Canadá, Suécia e Turquia, que contrariam este padrão.

Como se verifica, o declive da recta é positivo, o que traduz a associação positiva existente entre as variáveis em estudo.

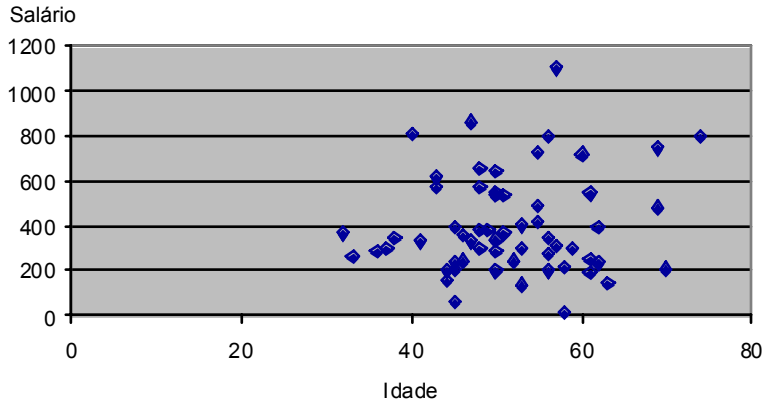
2.6.2. Exemplo 2 – Salários dos executivos (Fonte: lib.stat.cmu.edu)

A revista Forbes apresentou no dia 8 de Novembro de 1993 para as primeiras 59 empresas, da categoria de pequenas empresas, que apresentaram mais lucros nos últimos 5 anos, os salários (em milhares de dólares) e as idades dos administradores (Chief executive officer):

Idade	Salário	Idade	Salário	Idade	Salário	Idade	Salário	Idade	Salário	Idade	Salário
53	55	55	498	50	343	44	206	40	808	44	155
43	50	50	643	50	536	46	250	61	543	56	802
33	49	49	390	50	543	58	21	63	149	50	200
45	47	47	332	58	217	48	298	56	350	56	282
46	69	69	750	53	298	38	350	45	242	43	573

55	51	51	368	57	1103	74	800	61	198	48	388
41	48	48	659	53	406	60	726	70	213	52	250
55	62	62	234	61	254	32	370	59	296	62	396
36	45	45	396	47	862	51	536	57	317	48	572
45	37	37	300	56	204	50	291	69	482		

Pretende-se averiguar se se poderá admitir que os salários aumentam com a idade. Para ter uma ideia da associação entre a variável Idade e a variável Salário desenha-se o diagrama dos pontos de coordenadas (Idade, Salário) para as diferentes empresas consideradas:



A nuvem de pontos apresenta-se dispersa, sem nenhum padrão definido, dando a entender que não se pode admitir que quanto maior for a idade, maior será o salário auferido pelo executivo.

A existência de uma associação entre duas variáveis significará necessariamente uma relação de causa efeito?

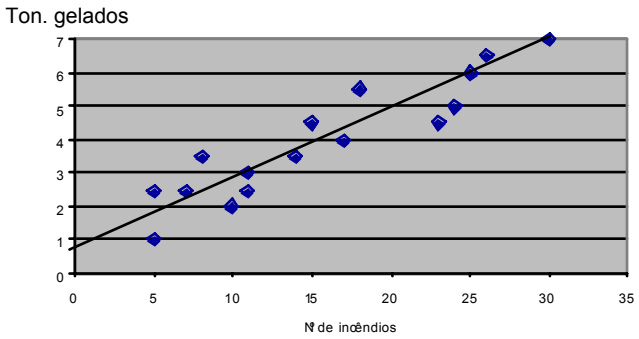
Se se detectar algum grau de associação entre duas variáveis, deve-se tomar cuidado com a interpretação que se dá a essa associação. Efectivamente, nem sempre a existência de associação entre duas variáveis significa uma relação de *causa efeito*. Pode haver outras variáveis, relacionadas com as variáveis em estudo, que façam com que se verifique essa associação, como se exemplifica a seguir.

2.6.3. Exemplo 3 – O consumo de gelados aumenta com o número de incêndios?

Registou-se durante o verão de vários anos, o número de incêndios que deflagraram e a quantidade (em toneladas) de gelados consumidos. Os resultados são apresentados na seguinte tabela: (Dados fictícios)

Nº incêndios	Ton. gelados	Nº incêndios	Ton. gelados	Nº incêndios	Ton. gelados
24	5	17	4	11	2,5
18	5,5	23	4,5	8	3,5
11	3	5	2,5	15	4,5
5	1	10	2	7	2,5
25	6	26	6,5		
14	3,5	30	7		

Construído o diagrama de dispersão, não deixa dúvidas o elevado grau de associação entre as variáveis nº de incêndios e quantidade de gelados consumidos.



Como se verifica através do gráfico, de um modo geral, quanto maior é o nº de incêndios, maior é a quantidade de gelados consumida. Será que tem algum sentido dizer que o consumo de gelados aumenta com o nº de incêndios? Obviamente que não. Neste caso facilmente se deduz que existe uma terceira variável, a intensidade de calor, que provoca um aumento das duas variáveis em estudo.

2.6.4. Exemplo 4 – Número de pessoas por aparelho de TV, tempo médio de vida

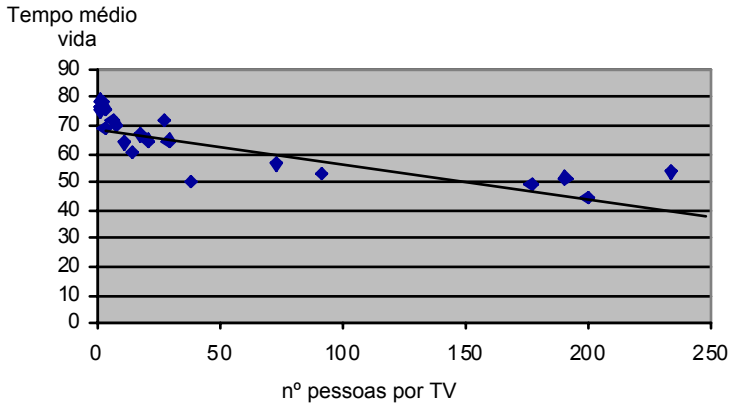
Diminuir o número de pessoas por aparelho de TV aumenta o tempo médio de vida? (Rossamn, 1995)

Para um conjunto de 22 países registou-se o número de pessoas por aparelho de TV, assim como o tempo médio de vida. Os resultados são apresentados na seguinte tabela:

País	Temp.méd.vida	Nº pes.porTV	País	Temp.méd.vida	Nº pes.porTV
Angola	44	200	México	72	6.6
Austrália	76.5	2	Marrocos	64.5	21
Cambodja	49.5	177	Paquistão	56.5	73
Canada	76.5	1.7	Rússia	69	3.2
China	70	8	África do Sul	64	11
Egipto	60.5	15	Sri Lanka	71.5	28
França	78	2.6	Uganda	51	191
Haiti	53.5	234	Reino Unido	76	3
Iraque	67	18	Est. Unidos	75.5	1.3
Japão	79	1.8	Vietnam	65	29
Madagascar	52.5	92	Yemen	50	38

A representação dos pontos de coordenadas (Nº de pessoas por aparelho de TV, Tempo médio de vida) num diagrama de dispersão, permite-nos concluir da existência de uma associação de uma certa intensidade, mas negativa, isto é, quanto menor for o número de pessoas por aparelho de TV, maior será o tempo médio de vida.

Poder-se-á então concluir que uma forma de aumentar o tempo médio de vida das populações é aumentando o número de aparelhos de TV, de forma a diminuir o racio (nº pessoas/nº aparelhos TV)?



2.7. Modelos de regressão linear

Objectivos a atingir:

- ✓ Ensinar a sumariar a relação linear existente entre duas variáveis, através de uma recta.
- ✓ Apresentar uma medida que além de indicar a força com que duas variáveis se associam linearmente, também dá indicação da "bondade" do ajustamento linear.

No módulo anterior em que se representaram graficamente conjuntos de pontos (x_i, y_i) num diagrama de pontos ou diagrama de dispersão, verificou-se que para alguns conjuntos

de pontos, se verificava a existência de uma certa associação linear traduzida pelo padrão da nuvem de pontos, na forma de uma oval, mais ou menos alongada. Pretende-se, nestes casos, introduzir um modelo matemático que traduza a relação entre os pontos, nomeadamente proceder a *um ajustamento de uma recta* a esses conjunto de pontos.

Recomenda-se que se comece por ajustar, a olho, uma recta de equação y=a+bx, que permita descrever como se reflectem em y – variável resposta, as modificações produzidas na variável x – variável explicativa, e que se determinem os coeficientes da recta a partir de 2 pontos escolhidos de forma conveniente.

Seguidamente falar-se-á na recta de regressão y = a+bx, cujos coeficientes são determinados de forma a minimizar a soma dos quadrados dos desvios $[y_i - (a+bx_i)]$. Uma vez determinados os coeficientes desta recta utilizando a máquina de calcular, sugere-se que se comparem as duas rectas – a ajustada empiricamente e a recta de regressão.

Sugere-se que sejam dadas as expressões que permitem calcular os coeficientes da recta de regressão, de onde se deduz que a recta de regressão passa pelo ponto $(\bar{x}, \bar{y})$. Esta propriedade poderá servir para obter uma recta ajustada a partir de dois pontos, em que um dos pontos é o de coordenadas $(\bar{x}, \bar{y})$.

Um processo visual simples de verificar se um ajustamento é razoável, é calcular os resíduos, isto é, para cada x_i, a diferença entre o valor dado y_i e o valor ajustado $\hat{y}_i = a+bx_{li}$ e representar num sistema de eixos coordenados os pontos $(x_i, y_i - \hat{y}_i)$. Se estes pontos se apresentarem aleatoriamente para cima e para baixo do eixo dos xx, sem um padrão bem definido, podemos esperar que os ajustamento seja bom. Este processo permite identificar os outliers como sendo os pontos que vão dar origem a grandes resíduos.

Utilizar a recta de regressão num dos seus objectivos fundamentais, isto é na predição de um valor para a variável resposta, a partir de um valor dado para a variável explicativa. Ao utilizar a

recta na predição tem que se ter o cuidado de definir à partida qual a variável explicativa e qual a variável resposta.

Devem ser referidas, nomeadamente dando exemplos, limitações da recta de regressão, quando existem outliers.

Posteriormente recomenda-se a definição do coeficiente de correlação, como uma medida que mede o maior ou menor grau de associação linear, com que as variáveis de associam. Deve ser apresentada a fórmula

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

que permite o seu cálculo e que deve ser utilizada para justificar graficamente o maior ou menor valor obtido para o coeficiente de correlação, conforme o aspecto da nuvem de pontos.

Devem ser enunciadas as propriedades do coeficiente de correlação, assim como devem ser realçadas as suas vantagens e desvantagens. Chamar a tenção para que não se deve calcular o coeficiente de correlação entre duas variáveis sem uma representação gráfica prévia dessas variáveis, que permita visualizar a existência de uma associação linear.

Devem ser referidas, nomeadamente dando exemplos, limitações do coeficiente de correlação, quando existem outliers.

Na interpretação do coeficiente de correlação deve-se chamar a atenção para o facto de que a existência de correlação elevada entre duas variáveis não significa necessariamente uma relação de causa-efeito. Pode verificar-se a existência de uma ou mais variáveis relacionadas com as variáveis em estudo, a provocar aquelas correlações referidas como correlações falsas.

Recomenda-se que se enuncie o seguinte resultado, que permite interpretar o coeficiente de correlação no contexto da recta de regressão: O quadrado do coeficiente de correlação mede a proporção da variabilidade na variável y, que é explicada pela relação linear entre y e x. Tendo em consideração que a recta de regressão é um modelo que se ajustou aos dados, melhor ou pior, o resultado anterior permite ajuizar dessa “bondade” e dar-nos uma indicação da confiança que devemos ter quando utilizamos a recta de regressão para fazer predições.

Deve ser ainda chamada a atenção para o perigo da utilização da recta de regressão para fazer extrapolações.

2.7.1. Exemplo 1 – Relação entre a altura e a idade de crianças

Os dados da tabela seguinte representam a idade (meses) e a altura (cm) das crianças de uma turma de uma escola privada.

Criança	Idade	Altura
1	109	137.6
2	113	147.8
3	115	136.8
4	116	140.7
5	120	145.4
6	126	148.5
7	129	148.3
8	130	147.5
9	133	148.8
10	135	148.7
11	137	152.0
12	139	150.6
13	142	149.9

a) Construa um diagrama de dispersão para os pontos (Idade, Altura).

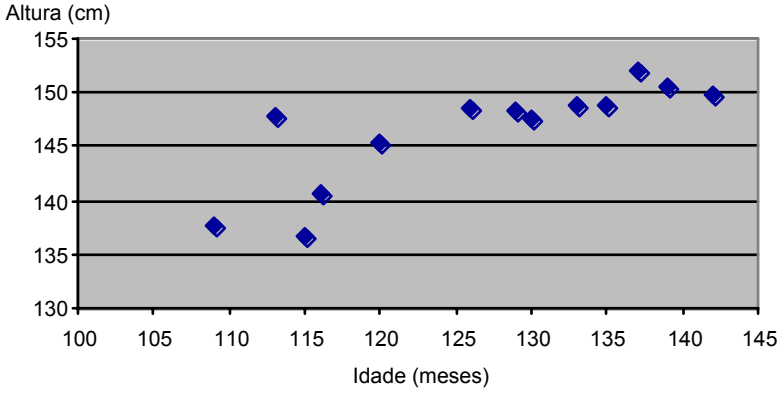
b) Tendo em conta o diagrama de dispersão, se achar conveniente ajuste uma recta aos dados, escolhendo dois pontos que ache convenientes.

c) Utilizando a máquina de calcular construa a recta de regressão da Altura sobre a Idade.

d) Qual a altura prevista para uma criança de 118 meses?

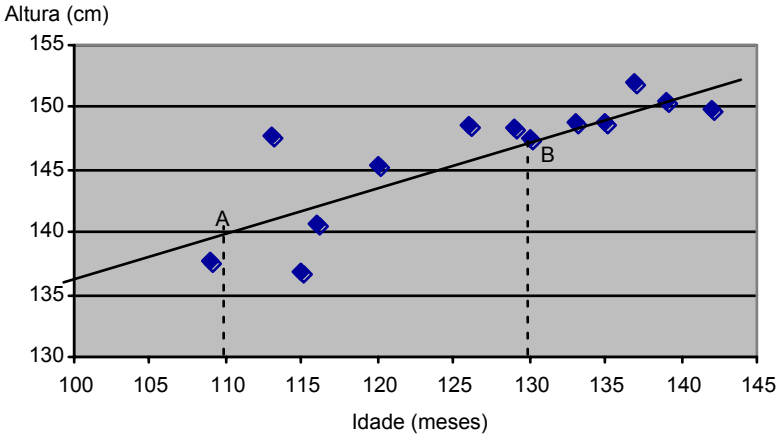
Resolução:

a) O diagrama de dispersão para os pontos (Idade, Altura), em que estamos a considerar a Idade como variável explicativa e a Altura como variável resposta, é:



A forma da nuvem de pontos sugere a existência de uma certa associação linear entre as variáveis.

b) Uma recta possível é a que se apresenta a seguir, e que foi ajustada “a olho”:



Considerámos os pontos A=(x_1, y_1) e B=(x_2, y_2) de coordenadas (110, 140) e (130, 147.5) para construir uma recta $\hat{y}=a+bx$, cujos coeficientes passamos a construir:

$$b = \frac{y_2 - y_1}{x_2 - x_1} = \frac{147.5 - 140}{130 - 110} = 0.375 \qquad \text{e}$$

$$a = y_1 - b \times x_1 = 140 - 0.375 \text{x} 110 \approx 99$$

Assim vem para equação da recta ajustada

$$\hat{y} = 99 + 0.375\, x$$

c) Utilizando a máquina de calcular obtivemos a recta de regressão

$$\hat{y} = 100 + 0.368\, x$$

cujos coeficientes não se distinguem muito dos obtidos para a recta ajustada a olho.

d) A partir da equação da recta de regressão obtemos que a altura prevista para uma criança com 118 meses seria 143.4 cm. Se tivéssemos utilizado a recta ajustada "a olho" para prever a altura de uma criança com a mesma idade, obteríamos o valor 142.3cm.

Observação: Representamos a recta ajustada por $\hat{y}$, para não confundir os valores ajustados $\hat{y}_i = a+bx_i$, com os valores dados y_i.

2.7.2. Exemplo 2 – O preço dos carros FIAT e a cilindrada

Considere os seguintes dados que dizem respeito à cilindrada e ao preço base de vários modelos de carros da marca FIAT (Valores de tabela de 1998):

Modelo	Cilindrada	Preço base	Modelo	Cilindrada	Preço base
Cinquecento S	899	1085	Punto Sport 16V	1342	1849
Cinquecento Soleil	899	1222	Punto 55 Star	1100	1493
Cinquecento Sport	899	1247	Bravo 1.4 S	1398	1865
Panda Jolly	899	1097	Bravo 1.6 SX	1598	2243
Panda 4x4	999	1481	Bravo TD 100 GT	1598	2308
Panda 4x4 C.Club	999	1640	Bravo 2.0 HGT	1998	2991
Punto 55 S	1100	1292	Brava 1.4 S	1398	1930
Punto 60 SX SEL	1100	1744	Brava TD 100 SX	1598	2202
Punto 75 SX	1242	1654	Marea 1.4 SX	1398	2215
Punto 85 16V ELX	1242	1701	Marea TD 100 ELX	1898	2605
Punto TD 70 ELX	1242	1507	Marea Weekend1.6ELX	1598	2649
Punto GT	1372	2285	Marea Weekend 2.0 H	1995	3576
Punto 60 S Cabrio	1242	2224	Coupe Fiat 1.8 16V	1795	3374
Punto 85 16V ELX Ca	1242	2577	Barchetta 1.8 16V	1795	3323

a) Construa um diagrama de dispersão considerando para variável explicativa a cilindrada e para variável resposta o preço base.

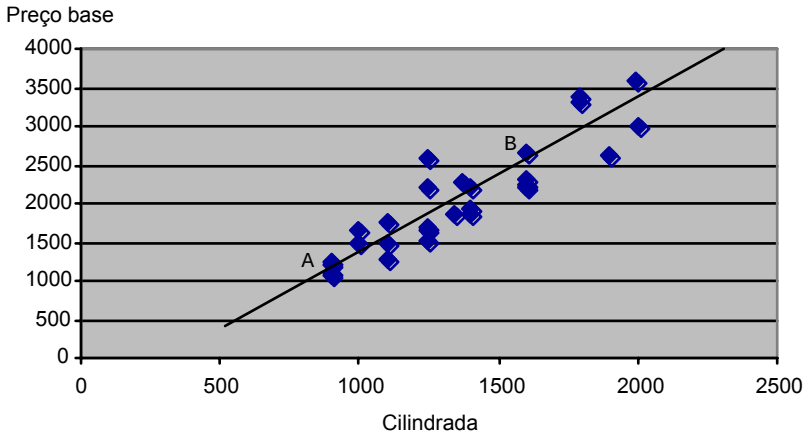
b) O diagrama de pontos revela alguma associação entre a cilindrada e o preço base? Caso afirmativo comente o tipo e o grau de associação.

c) Apesar das conclusões que chegou na alínea anterior, será possível existirem carros com maior cilindrada do que outros, mas com preço base inferior?

d) Obtenha uma recta ajustada e prediga o valor do preço base do modelo Bravo 1.4 SX, que tem de cilindrada 1398.

Resolução:

a) O diagrama de pontos tem o seguinte aspecto



b) O diagrama revela a existência de uma associação linear positiva entre a cilindrada e o preço base do modelo de carro. A associação linear positiva significa que em média, quando a cilindrada aumenta, aumenta também o preço base do modelo.

c) Sim, já que, como dissemos na alínea anterior, quando a cilindrada aumenta, em média o preço aumenta. Isto significa que existem pontos em que a variável explicativa varia no sentido inverso da variável resposta. São exemplos os pares de pontos (999, 1640) e (1100, 1292).

d) Para construir a recta ajustada considerámos os pontos A(899, 1200) e B(1598,2649), obtendo-se a recta ajustada

$$\ddot{y} = -551 + 2x$$

O valor predito para o preço base do modelo Bravo 1.4 SX é obtido substituindo o x por 1398

$$\ddot{y} = -551 + 2x1398 = 2245$$

ou seja, o preço base predito é de 2245 contos.

2.7.3. Exemplo 3 (Turkman, 1997) – Apanha automática de uvas

As vinhas estão geralmente dispostas de uma maneira muito regular, com longas filas de videiras dispostas paralelamente e separadas por um estreito arruamento. Isto permite que máquinas automáticas passem pelos arruamentos para a apanha da uva, que é feita por um braço rotativo. De modo a estudar a eficiência da máquina, registou-se o número de cachos não retirados, fazendo variar a velocidade de rotação do braço, enquanto a máquina viajava através do arruamento a uma velocidade constante. O resultado da experiência encontra-se na tabela seguinte:

Proporção de cachos não apanhados - y	Velocidade de motor (r.p.m) - x	Proporção de cachos não apanhados - y	Velocidade de motor (r.p.m) - x
0.100	3.16	0.034	4.16
0.067	3.16	0.026	4.16
0.168	3.16	0.016	4.16
0.132	3.16	0.008	4.16
0.051	3.66	0.009	4.66
0.093	3.66	0.014	4.66
0.027	3.66	0.002	4.66
0.025	3.66	0.003	4.66

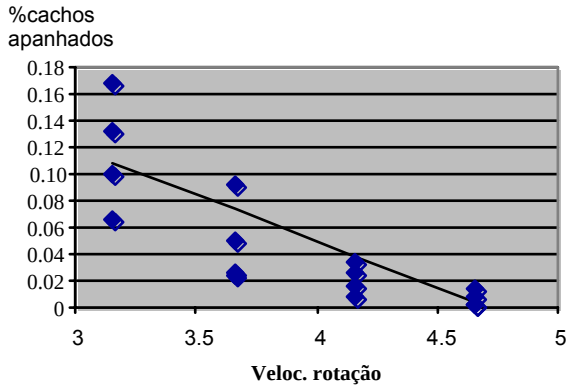
Pretende-se averiguar de que modo a velocidade do motor afecta a proporção de cachos não apanhados, para poder decidir, por exemplo, a velocidade adequada.

a) Represente os dados num diagrama de pontos e comente a representação obtida.

b) Caso na alínea anterior tenha concluído pela existência de uma associação linear entre os dados, encontre um modelo conveniente.

Resolução:

a) Considerando como variável explicativa a velocidade de rotação e como variável resposta a proporção de cachos não apanhados, obtém-se a seguinte representação gráfica:



b) Neste exemplo não é imediato que o melhor ajustamento seja o linear. No entanto optamos, mesmo assim, por ajustar uma recta, que se encontra representada no gráfico anterior e cuja equação é $\hat{y} = 0.33 - 0.07x$.

2.7.4. Exemplo 4 (Chatterjee, 1995) – Adopção internacional de crianças

Na tabela seguinte apresentamos o número de vistos passados pelos serviços de Emigração e Naturalização dos EUA, com vista à adopção de crianças estrangeiras, pelas famílias americanas. Os dados referem-se aos anos de 1988 e 1991.

	País	1988	1991		País	1988	1991
1	Africa	28	41	21	Jamaica	30	39
2	Belize	6	4	22	Japan	69	83
3	Bolivia	21	51	23	Lebanon	23	17
4	Brazil	164	178	24	Mexico	123	106
5	Cambodia	0	59	25	Nicaragua	5	11
6	Canada	12	12	26	Oceania	15	16
7	Chile	252	263	27	Pakistan	10	9
8	China	52	62	28	Panama	23	10
9	Colombia	699	527	29	Paraguay	300	177
10	Costa Rica	73	55	30	Peru	142	722
11	Dominican Rep	54	50	31	Phillipines	476	417
12	Ecuador	41	11	32	Poland	51	95
13	El Salvador	88	122	33	Portugal	17	10
14	Greece	10	5	34	Romania	0	2552
15	Guatamala	209	324	35	South Korea	4942	1817
16	Haiti	41	52	36	Taiwan	56	55
17	Honduras	161	244	37	Thailand	75	127
18	Hong Kong	49	40	38	Turkey	11	6
19	Hungary	6	25	39	Vietnam	1	17
20	India	698	448				

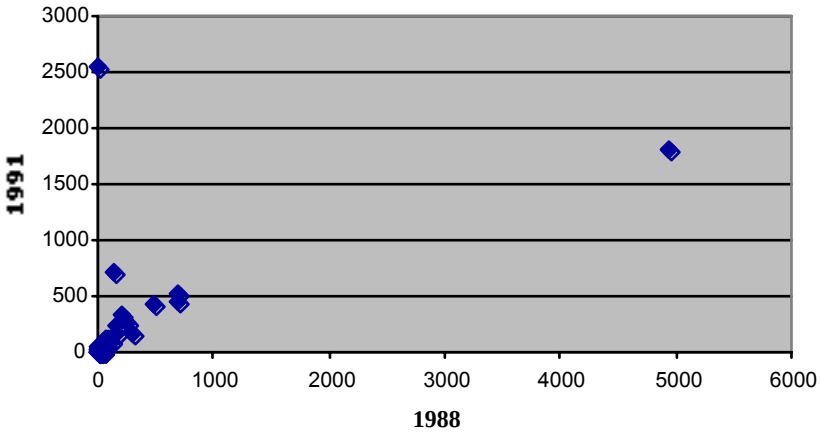
a) Represente os dados num diagrama de pontos e comente a representação obtida.

b) Retire os pontos que lhe pareçam outliers e ajuste uma recta aos restantes.

c) Represente graficamente os resíduos.

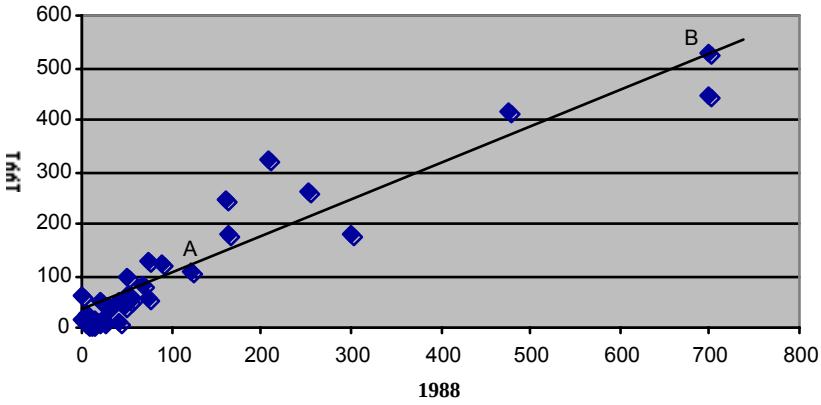
Resolução:

a) O diagrama dos pontos correspondentes ao número de vistos passados em 1988 – variável explicativa (x) e ao número de vistos passados em 1991 – variável resposta (y), para as diferentes regiões consideradas, tem o seguinte aspecto:



Assinalámos 3 pontos correspondentes ao Peru, Romania e S. Korea, que considerámos como outliers, uma vez que os seus valores saem fora do contexto dos restantes. Aparentemente os outros pontos parecem seguir um padrão linear.

b) Retirando os 3 pontos assinalados anteriormente, obtemos o seguinte diagrama de pontos, onde representámos uma recta que parece ser um "bom" ajustamento.



Considerámos os pontos A(123,110) e B(699, 527) para determinar os coeficientes a e b da recta ajustada $\hat{y}=a+bx$. O ponto A não é nenhum dos pontos dados, pois embora a sua abcissa coincida com o valor de 1988 para o México, a sua ordenada é ligeiramente superior a 106, valor correspondente para 1991, pelo que considerámos 110.

Os coeficientes da recta ajustada são

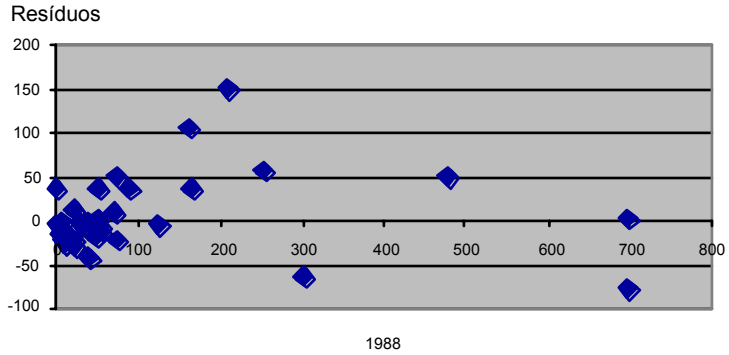
$$b = \frac{527 - 110}{699 - 123} = 0.72 \qquad \text{e} \qquad a = 110 - 0.72 \times 123 = 21.44$$

pelo que a equação da recta é

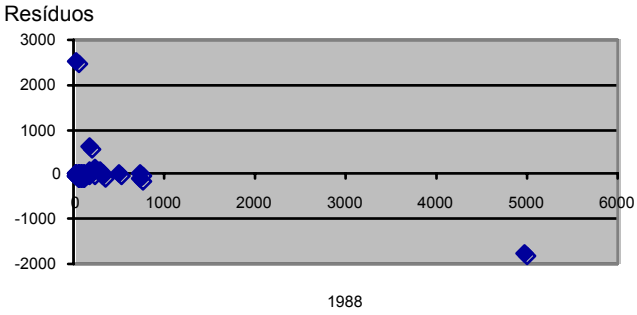
$$\hat{y} = 21.44 + 0.72\, x$$

O modelo anterior sugere que se possa utilizar o número de vistos passados num ano, para predizer o número de vistos passados noutro ano. Tem no entanto de se ter cuidado com o seguinte: se o país em estudo mudar a política de adopção, as predições não têm qualquer valor, já que o modelo deixa se se aplicar. Aliás, um dos pontos inicialmente considerados e que resolvemos retirar do estudo por se considerar um outlier tem uma explicação: a seguir ao derrube do regime comunista a Romania incentivou a adopção internacional, mas rapidamente voltou atrás como se poderia constatar se se analisassem os dados correspondentes a 1992.

c) Pode-se verificar que um ajustamento é bom, calculando os resíduos, isto é, as diferenças entre os valores ajustados, obtidos para a variável resposta, utilizando a recta ajustada, e os valores observados. Para obter os resíduos , substitui-se cada valor x_i da variável explicativa na recta ajustada, obtendo-se o valor ajustado $\hat{y}_i$, que se subtrai do valor dado y_i. Para o caso em estudo a representação gráfica dos resíduos tem o seguinte aspecto

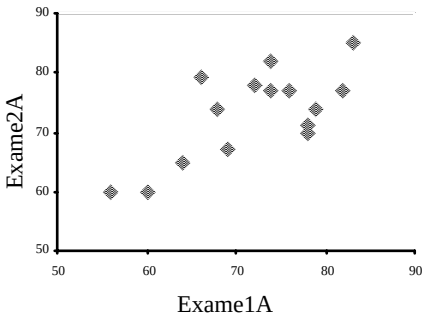


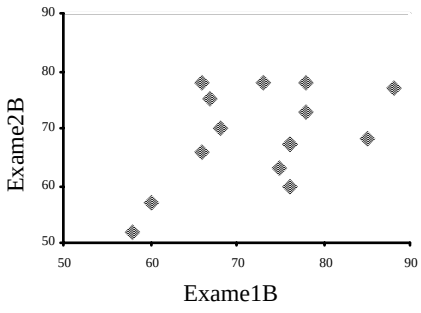
Os pontos apresentam-se aleatoriamente para um e outros lado do eixo dos xx, o que é sintoma de que com o ajustamento se conseguiu obter a tendência da associação entre os dados. Verifica-se ainda existirem 2 pontos, com resíduos elevados, correspondentes à Guatemala e às Honduras. Chama-se a atenção que não estão representados os valores correspondentes ao Peru, Romania e S. Korea, por à partida os termos considerados outliers. Vejamos qual a representação obtida para os resíduos, considerando todos os pontos e a mesma recta ajustada:

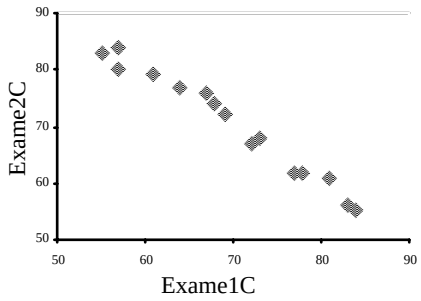


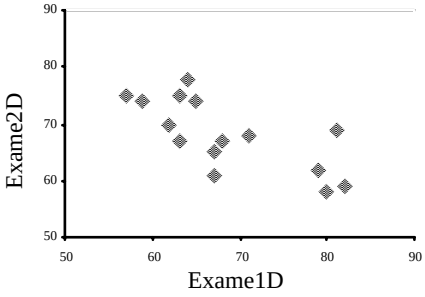
2.7.5. Exemplo 5 (Rossman, 1996) – Comparação de exames

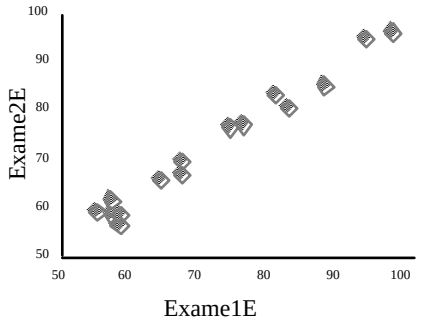
Considere os seguintes diagramas de dispersão correspondentes aos resultados de 2 exames de 6 classes (A-F).

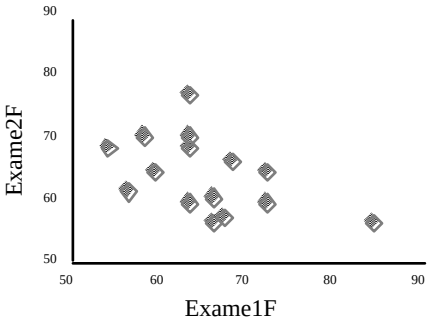












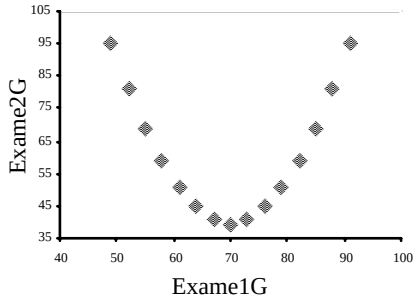
a) Preencha a seguinte tabela a partir da visualização dos gráficos

	Forte	Moderada	Fraca
Positiva			
Negativa			

b) Verifique se a tabela obtida na alínea anterior está consistente com os resultados da seguinte tabela:

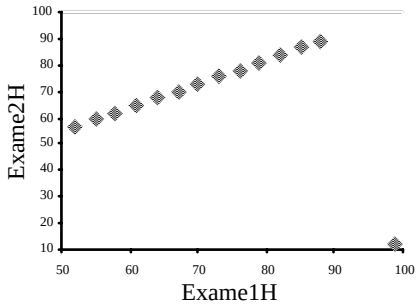
Classe	Correlação
A	0.71
B	0.47
C	-0.99
D	-0.72
E	0.99
F	-0.47

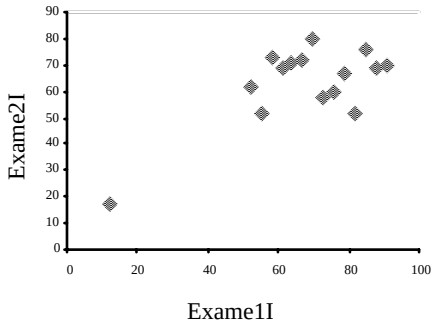
c) Considere agora a seguinte representação correspondente aos dados de uma classe G:



Como se verifica, existe uma forte associação entre os valores do exame 1 e os valores do exame 2. *Surpreendentemente* ao calcular o coeficiente de correlação obtemos o valor 0! Comente.

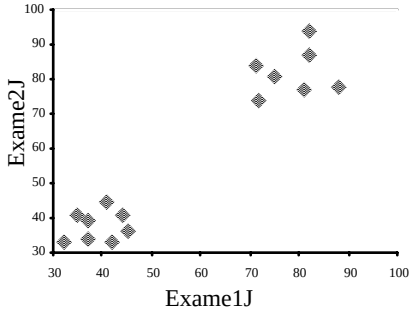
d) Considere agora as duas representações correspondentes às notas obtidas pelas classes H e I:





O valor para o coeficiente de correlação é respectivamente 0.04 e 0.70 para as classes H e I, o que continua a ser surpreendente! Repare-se que relativamente à classe H todos os pares menos 1 seguem um padrão linear, tendo-se obtido para o coeficiente de correlação um valor próximo de zero, enquanto que para a classe I, em que os valores se apresentam mais ou menos dispersos, obtivemos um valor relativamente alto. No entanto, se retirarmos a cada um dos conjuntos de dados anteriores o "outlier", já o valor do coeficiente de correlação passa para 0.9997 e 0.13, respectivamente para as classes H e I. Comente os resultados anteriores.

e) Finalmente consideremos o seguinte diagrama de dispersão correspondente à classe J:



Da análise da representação anterior verificamos existirem dois grupos distintos de alunos: uns muito bons e outros muito maus. Embora para cada um dos grupos se verifique uma ligeira tendência para uma associação positiva, o facto é que o valor do coeficiente de correlação é 0.95, bem superior ao valor que seria de esperar. Comente.

Resolução:

a) A visualização dos gráficos anteriores leva-nos a supor que entre os dois exames se possa admitir o seguinte tipo de associação:

b)

	Forte	Moderada	Fraca
Positiva	E	A	B
Negativa	C	D	F

c) Sim, pois está de acordo com o facto de se ter referido o tipo de associação como positiva ou negativa e ainda o grau de associação como forte, moderada ou fraca.

d) Não é assim tão surpreendente se nos lembrarmos que o que o coeficiente de correlação mede é o grau de associação linear e não outro tipo de associação, como a associação curvilínea, presente nos dados da representação anterior.

e) O exemplo que acabámos de dar mostra que o coeficiente de correlação não é uma medida *resistente*, já que é muito influenciado pelos ”outliers”. Este facto não é de estranhar, já que no cálculo do coeficiente de correlação entramos com a média, que já vimos ser uma medida não resistente.

f) Os exemplos que acabámos de ver, elucidam-nos sobre as limitações do coeficiente de correlação como medida de associação entre duas variáveis. Pode ser perigoso apresentar o coeficiente de correlação como uma medida de associação entre duas variáveis, sem primeiro ter feito a representação gráfica dos pares de valores das variáveis.

2.7.6. Exemplo 6 – Número de pessoas por aparelho Tv e tempo médio de vida

(Continuação do exercício do módulo anterior)

Para os dados do exemplo 4 do módulo anterior, calcule o coeficiente de correlação e comente os resultados obtidos.

Resolução: O coeficiente de correlação é igual a –0.80. Este valor indica uma forte associação (linear, indicada pela representação gráfica) negativa entre o número de pessoas por aparelho de TV e o tempo médio de vida ou seja, quanto maior for o número de pessoas por aparelho de TV, menor é o tempo médio de vida. Será que então se pode aumentar o tempo médio de vida da população de um país, aumentando o número de aparelhos de TV? Seria ridículo pensar desta maneira, pois este é um exemplo em que sobressai que não se pode admitir uma relação de *causa-efeito*. Obviamente existem outras variáveis não observadas –*variáveis perturbadoras* - relacionadas com o nível de vida na população, que provocam alterações nas duas variáveis que estamos a estudar e que explicam a forte correlação verificada.

2.7.7. Exemplo 7 – Nos casais existe alguma relação entre a altura do homem e da mulher?

Pensa-se que os casais têm tendência para terem alturas semelhantes. Considere os seguintes pares que dizem respeito às alturas (em cm) de 10 casais:

Mulher	170	164	167	165	164	165	166	165	162	163
Homem	183	168	178	173	167	164	165	170	165	164

a) Represente os pontos num diagrama de dispersão. Tendo em consideração a representação

gráfica obtida, espera que o coeficiente de correlação seja grande ou pequeno? Perto de 1 ou de –1?

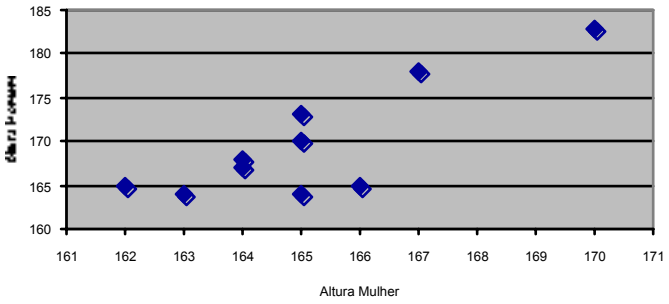
b) Calcule o valor do coeficiente de correlação.

c) Adicione 10 às alturas das mulheres? Qual o valor do coeficiente de correlação?

d) Se todas as mulheres escolherem um homem mais alto do que elas 5 cm, qual será o coeficiente de correlação?

Resolução:

a)



b) A representação gráfica dos pontos sugere que o coeficiente de correlação seja razoavelmente grande, perto de 1.

c) coeficiente de correlação é igual a 0.83. Calculámos o seu valor a partir da expressão

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

onde se representa por x uma variável e por y a outra variável.

d) O valor obtido para o coeficiente de correlação não se altera. Efectivamente se adicionarmos o mesmo valor a todos os elementos de uma das variáveis, o tipo e o grau da associação linear não se altera.

e) O coeficiente do correlação será 1, pois as variáveis estão relacionadas por uma relação determinística, isto é,

Altura homem = Altura mulher + 5

Assim, dados dois pares quaisquer, sempre que ao passar de um para o outro se aumenta (diminui) a altura da mulher, também aumenta (diminui) a altura do homem, existindo assim uma concordância perfeita.

2.7.8. Exemplo 8 (Murteira, 1993) – Colheita e preço do vinho

Considere as seguintes observações referentes às colheitas de vinho tinto no continente (em milhares de hectolitros) e ao preço por litro, no período de 1944 a 1961:

Ano	Colheita	Preço	Ano	Colheita	Preço
1944	10939	1.52	1953	8790	2.02
1945	7826	2.29	1954	8959	1.68
1946	7165	3.54	1955	8289	1.78
1947	7807	2.09	1956	7910	1.99
1948	6028	2.46	1957	6775	3.02
1949	6037	2.5	1958	6088	3.23
1950	6458	2.62	1959	6381	2.98
1951	6981	2.57	1960	8600	2.74
1952	4233	3.38	1961	4805	3.38

a) Construa o diagrama de dispersão e tente ajustar uma tendência conveniente. Interprete o coeficiente b da recta de regressão.

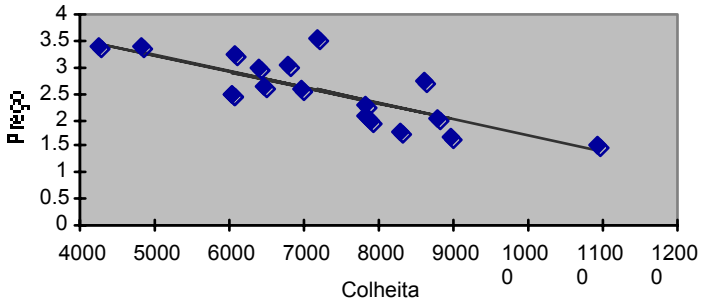
b) Determine a correlação entre as variáveis "colheita" e "preço". Comente.

c) Passe a variável "colheita" para decalitros e volte a fazer a alínea anterior. Comente.

d) Tendo em consideração o resultado obtido na alínea b) o ajustamento considerado em a) parece-lhe razoável?

Resolução:

a) O diagrama de dispersão dos pontos (Colheita, Preço) tem o seguinte aspecto:



Da representação anterior verificamos que em média o aumento da colheita provoca uma diminuição do preço. A tendência da relação entre a colheita e o preço é dada pela recta de regressão, cuja equação y = a+bx, é:

$$y = 4.73 - 0.0003x$$

Interpretação do coeficiente b= –0.0003 : Um acréscimo de 10^6 hl na colheita, implica um decréscimo, em média de 0.30 escudos no preço por litro.

Interpretação do coeficiente a=4.73: O valor predito para o preço do litro, quando a colheita fosse nula, seria de 4.73.

b) O coeficiente de correlação r entre as variáveis Colheita e Preço, calculado na máquina de calcular chamando a função respectiva ou através da equação

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

onde se representa por x uma variável e por y a outra variável, é igual a r = -0.784. Este valor traduz a existência de uma forte associação negativa entre a Colheita e o Preço do litro.

c) Mudando de unidades, obtém-se o mesmo valor para o coeficiente de correlação. Aliás esta propriedade poderia ser deduzida a partir da expressão do coeficiente de correlação, em que se verifica que este é independente das unidades das variáveis x e y.

d) O quadrado do coeficiente de correlação dá a percentagem da variabilidade na variável resposta que é explicada pela recta de regressão. Assim temos que no caso em estudo 61.5% das variações no preço do vinho por litro, é explicada pelas variações na Colheita.

2.8. Relação entre variáveis qualitativas

Objectivos a atingir:

- ✓ Apresentar um modo eficaz de organizar informação de tipo qualitativo.
- ✓ Chamar a atenção para a utilização incorrecta que, por vezes, se faz da leitura de percentagens a partir de tabelas.

No módulo anterior foram exploradas as relações entre variáveis de tipo quantitativo. Pretende-se neste módulo estudar algumas formas de explorar as relações entre variáveis de tipo qualitativo. Chama-se a atenção para o facto de que as variáveis envolvidas podem ser por inerência de tipo qualitativo (sexo, idade, etc), enquanto que outras foram categorizadas por se ter procedido a agrupamentos de variáveis de tipo quantitativo (idade, altura, etc).

O instrumento básico para a análise de dados bivariados, de tipo qualitativo é a representação dos dados em *tabelas de contingência*, cuja análise se faz calculando percentagens adequadas.

Na análise das tabelas de contingência devem ser referidas as *distribuições marginais* e as *distribuições condicionais*.

Finalmente sugere-se a utilização de *gráficos de barras segmentadas*, frequentemente utilizados nos meios de comunicação social, para representar as distribuições condicionais.

Devem-se referir ainda alguns cuidados a ter na análise de tabelas de contingência, referindo o paradoxo de Simpson, que refere a mudança possível de direcção de uma comparação ou associação, quando dados de vários grupos são combinados num único grupo.

2.8.1. Exemplo 1- Estado civil e categoria dos docentes

Suponha que uma universidade decidiu estudar o seu corpo docente quanto ao estado civil e categoria profissional, tendo obtido os seguintes resultados:

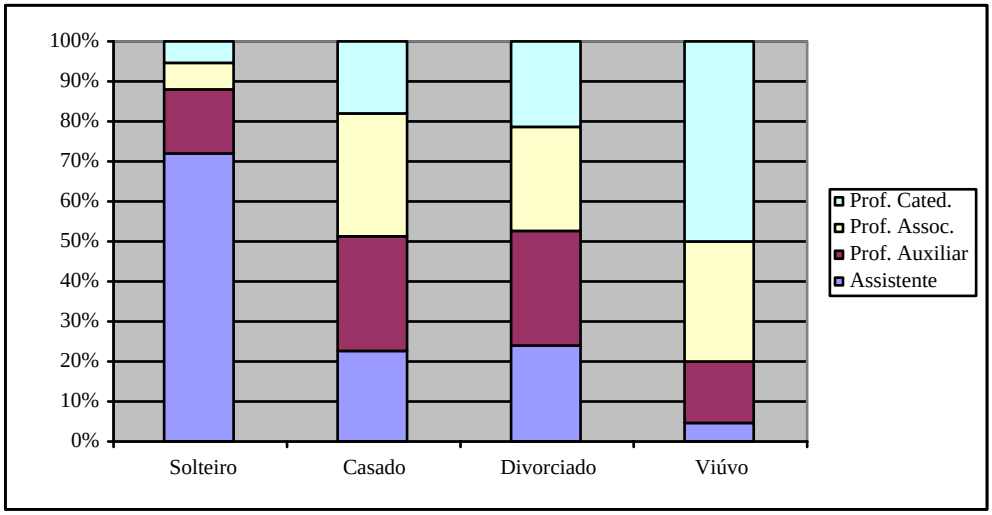
Estado civil Categoria	Solteiro	Casado	Divorciado	Viúvo	Total
Assistente	111	43	10	1	165
Prof. Auxiliar	25	54	12	3	94
Prof. Associado	10	58	11	6	85
Prof. Catedrático	8	34	9	10	61
Total	154	189	42	20	405

Na última coluna do lado direito apresentamos os totais de linha, que corresponde à *distribuição* da variável "categoria profissional". Analogamente, na última linha estão apresentados os totais de coluna, que correspondem à *distribuição* da variável "estado civil". A estas distribuições chamamos *distribuições marginais* (precisamente por se apresentarem nas margens da tabela!). Estas distribuições apresentadas separadamente não nos dão informação sobre a associação entre as variáveis em estudo. Tão pouco essa informação pode ser dada pelo diagrama de dispersão ou pela correlação.

Uma forma de descrever a relação entre variáveis qualitativas é através do cálculo de percentagens convenientes. Consideremos a tabela seguinte, obtida a partir da tabela anterior, dividindo o valor de cada célula pelo total de coluna correspondente:

Estado civil Categoria	Solteiro	Casado	Divorciado	Viúvo	
Assistente	0.721	0.228	0.238	0.050	0.407
Prof. Auxiliar	0.162	0.285	0.286	0.150	0.232
Prof. Associado	0.065	0.307	0.262	0.300	0.210
Prof. Catedrático	0.052	0.180	0.214	0.500	0.151
Total	1.000	1.000	1.000	1.000	1.000

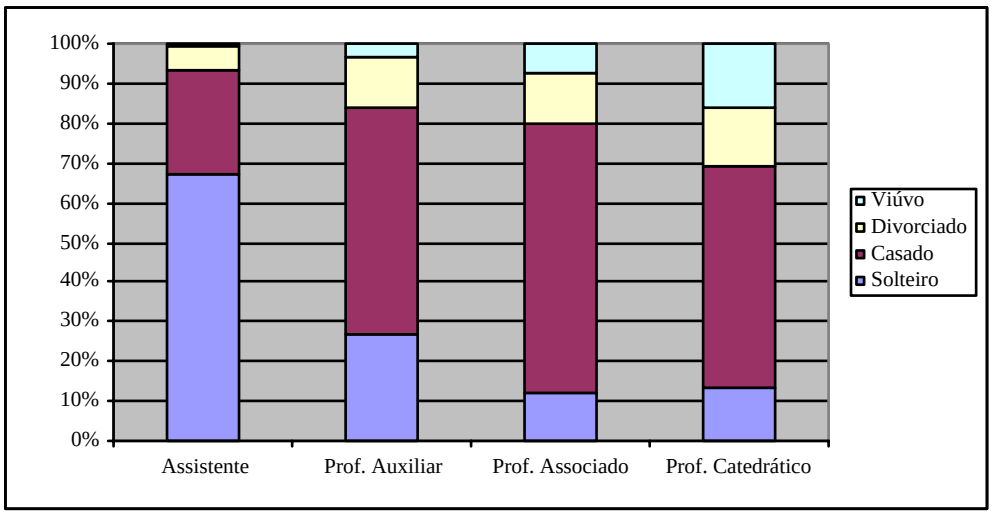
Nesta tabela apresentamos as *distribuições condicionais* da variável categoria profissional, relativamente às classes da outra variável estado civil. Temos assim que, por exemplo, nos solteiros a percentagem de assistentes é de aproximadamente 72%, enquanto que nos casados é de aproximadamente 23%. Estas distribuições condicionais podem ser visualizadas graficamente num diagrama de barras por segmentos, como se apresenta a seguir:



Se estivéssemos interessados nas distribuições condicionais da variável estado civil, condicional à variável categoria profissional, então a tabela a construir seria:

Estado civil Categoria	Solteiro	Casado	Divorciado	Viúvo	Total
Assistente	0.673	0.261	0.061	0.006	1.001
Prof. Auxiliar	0.266	0.574	0.128	0.032	1.000
Prof. Associado	0.118	0.682	0.129	0.071	1.000
Prof. Catedrát.	0.131	0.557	0.148	0.164	1.000
	0.380	0.467	0.104	0.049	1.000

A leitura que se deve fazer desta tabela é semelhante à que se fez da tabela anterior, mas tendo em atenção que agora a variável que está a condicionar é a categoria profissional. Por exemplo pode obter-se a informação de que aproximadamente 67% dos assistentes são solteiros, enquanto que casados são cerca de 26%. O diagrama de barras por segmentos correspondente a estas distribuições marginais tem o seguinte aspecto:



Podemos finalmente estar interessados na *distribuição conjunta* das duas variáveis, e então em vez de recolher a informação a partir da primeira tabela constrói-se uma outra em que a

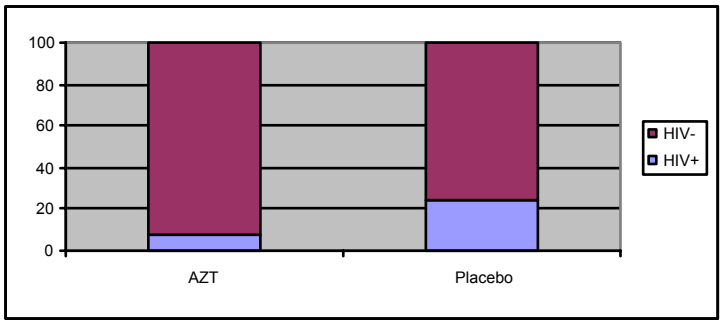
frequência absoluta de cada célula é substituída pela frequência relativa, relativamente ao total de docentes, pois as frequências relativas são mais fáceis de comparar:

Estado civil Categoria	Solteiro	Casado	Divorciado	Viúvo	Total
Assistente	0.274	0.106	0.025	0.002	0.407
Prof. Auxiliar	0.062	0.133	0.030	0.007	0.232
Prof. Associado	0.025	0.143	0.027	0.015	0.210
Prof. Catedrático	0.020	0.084	0.022	0.025	0.151
Total	0.380	0.467	0.104	0.049	1.000

Desta tabela imediatamente se conclui que, do pessoal docente, 3% são Professores Auxiliares e casados, enquanto que Assistentes e solteiros são mais de 27%.

2.8.2. Exemplo 2 (Rossman, 1996) – O vírus HIV e o medicamento AZT

O Newsweek de 7 de Março de 1994, relata uma experimentação realizada com 164 mulheres grávidas, positivas para o vírus HIV, que foram seleccionadas aleatoriamente para tomarem o medicamento AZT, durante a gravidez, enquanto que 160 mulheres foram seleccionadas aleatoriamente para um grupo de controlo, que tomou um placebo. No gráfico seguinte apresentam-se os resultados relativamente à presença do vírus, nos filhos das mulheres dos dois grupos:



a) A partir do gráfico estime a proporção de bebés HIV+, tanto para as mulheres que tomaram o AZT, como para as que tomaram o placebo.
b) Os resultados obtidos da experimentação foram de 13 crianças HIV+ para as mulheres que tomaram AZT, enquanto que para as mulheres do grupo de controlo o número de crianças HIV+ foi de 40. Calcule as verdadeiras percentagens e compare-as com as estimativas consideradas na alínea a).
c) Comente os resultados obtidos e diga se as diferenças para os dois grupos parecem importantes.

Resolução:

a) A percentagem de crianças HIV+ para o grupo AZT é aproximadamente 7%, enquanto que para o grupo de controlo é de aproximadamente 25%.

b) Percentagem de crianças HIV+ para o grupo AZT= $100x\frac{13}{164}$ = 8%

Percentagem de crianças HIV+ para o grupo controlo = $100x\frac{40}{160}$ = 25%

c) A percentagem de bebés HIV+ é mais do que três vezes superior no grupo que não tomou AZT. Os resultados parecem evidenciar o efeito de prevenção do medicamento AZT.

2.8.3. Exemplo3 (Moore, 1993) – Discriminação sexual nos candidatos a uma Universidade

A Upper Wabash Tech tem duas faculdades: de Gestão e de Direito. A seguir apresenta-se uma tabela de candidatos a essas faculdades, discriminados por sexo, faculdade e decisão de admissão.

	Gestão			Direito	
	Admitidos	Não admit.		Admitidos	Não admit.
Homens	480	120	Homens	10	90
Mulheres	180	20	Mulheres	100	200

a) Construa uma tabela de dupla entrada onde considera o sexo e o número de admitidos e não admitidos conjuntamente para as duas faculdades.
b) Calcule a percentagem de homens e mulheres que foram admitidas. Comente .
c) Calcule separadamente a percentagem de homens e mulheres que foram admitidos, nas duas faculdades. Comente.
d) Explique como é possível que aparentemente a Upper Wabash favoreça os homens, quando cada faculdade individualmente favorece as mulheres.

Resolução:

a)

	Admitidos	Não admitidos
Homens	490	210
Mulheres	280	220

b) Percentagem de homens admitidos = $100x\frac{490}{700} = 70\%$

Percentagem de mulheres admitidas = $100x\frac{280}{500} = 56\%$

Verifica-se que a percentagem de homens admitidos é substancialmente superior à percentagem de mulheres admitidas. Haverá discriminação contra as mulheres?

c) Percentagem de homens admitidos em Gestão = $100x\frac{480}{600} = 80\%$

Percentagem de mulheres admitidos em Gestão = $100x\frac{180}{200} = 90\%$

Percentagem de homens admitidos em Direito = $100x\frac{10}{100} = 10\%$

Percentagem de mulheres admitidos em Direito = $100x\frac{100}{300}$ = 33%

Os resultados anteriores permitem concluir que a percentagem de mulheres admitidas é superior à percentagem de homens admitidos, para as duas faculdades. Afinal, a haver discriminação, será contra os homens!

d) O paradoxo é devido ao facto de a maior dos homens se terem candidatado à faculdade de Gestão, onde é mais fácil de entrar.

3. Modelos de Probabilidade

3.1. Fenómenos aleatórios

Objectivos a atingir:

- ✓ Dar a entender aos alunos a diferença entre fenómeno determinístico e fenómeno aleatório.
- ✓ Alertar para as vantagens em encontrar modelos matemáticos apropriados para este tipo de fenómenos.

A existência de fenómenos que, por razões diversas, não são passíveis de ser descritos por leis determinísticas é a grande motivação para o aparecimento de modelos de probabilidade. Neste módulo sugerimos que se comece por dar exemplos de fenómenos físicos determinísticos (queda de um grave, movimento de um pêndulo,...) em contraponto com fenómenos que se podem considerar aleatórios devido à grande complexidade das leis físicas subjacentes (movimento de um dado ao ser lançado, movimento das partículas numa nuvem de pó, temperatura máxima observada numa data futura,...). Propõe-se ainda que se analise logo com algum detalhe o caso simples do lançamento de um dado, mas que só após o módulo seguinte se apresente a definição de fenómeno aleatório.

Exemplo 1

A face que fica virada para cima ao lançar um dado depende obviamente da sua posição inicial e de todo o movimento que ele descreve até se imobilizar. Tivéssemos nós acesso à lei desse movimento e saberíamos exactamente qual a face que iria ficar virada para cima em cada lançamento. Acontece que a expressão matemática dessa lei depende de muitos factores (do impulso inicial, da zona do dado que toca primeiro na superfície de contacto, das eventuais irregularidades dessa superfície de contacto, etc.). Em termos práticos todas estas condicionantes fazem com que se torne impossível saber, à partida, qual a face do dado que irá ficar virada para cima após cada lançamento.

No entanto, se admitirmos que o dado é composto de uma matéria homogénea não temos qualquer motivo para acreditar mais na saída de uma das faces em detrimento de outra. Podemos traduzir esta “crença” no seguinte modelo probabilístico:

Nº de pintas da face que fica virada para cima	1	2	3	4	5	6
Probabilidade	1/6	1/6	1/6	1/6	1/6	1/6

O facto de se admitir este modelo de probabilidade para o nº de pintas da face que fica virada para cima ao lançar um dado permite-nos agora construir modelos para experiências mais elaboradas, envolvendo vários lançamentos de um dado, ou o lançamento de vários dados.

Os modelos probabilísticos (ou modelos de probabilidade) são modelos matemáticos utilizados na representação e interpretação de fenómenos que, ou por serem demasiado complexos ou por terem um mecanismo de funcionamento desconhecido, não conseguem ser descritos por leis determinísticas.

3.2. Ex. de modelos de probabilidade em situação de simetria. Regra de Laplace.

Objectivos a atingir:

- ✓ Construir modelos de probabilidade para situações simples em que se admita como razoável o pressuposto de simetria ou equilíbrio.
- ✓ Calcular a probabilidade de alguns acontecimentos a partir dos modelos construídos.
- ✓ Construir modelos de probabilidade para situações um pouco mais complexas utilizando a regra do produto.

Em exemplos ligados aos chamados jogos de azar é quase sempre possível encontrar um espaço de resultados para cujos elementos, à partida, não se tem razão para admitir que não tenham igual probabilidade de ocorrer.

Começando por modelar os resultados de experiências muito simples (como o nº de pintas da face que fica virada para cima ao lançar um dado ou o naipe a que pertence uma carta extraída de um baralho) é possível introduzir a noção de **acontecimento** como um subconjunto do espaço de resultados e, explicando a razão de ser da regra do produto (com a modelação dos resultados de dois lançamentos de um dado, por exemplo), construir modelos para fenómenos aleatórios cujos resultados já não são equiprováveis (como na soma das pintas em dois lançamentos de um dado equilibrado).

Não se justifica, nesta disciplina, o estudo de modelos para situações que obriguem a utilizar técnicas de contagem. Em contrapartida devem ser dados exemplos ligados a experiências de amostragem, isto é, onde interesse encontrar um modelo de probabilidade para o valor observado de certa característica quando se "retira" ao acaso um ou mais indivíduos de uma população.

Este módulo deve ser finalizado com a apresentação e discussão com os alunos de alguns exemplos de fenómenos aleatórios para os quais não faça sentido utilizar argumentos de simetria, tentando ao mesmo tempo que eles se apercebam da necessidade de uma boa recolha de informação e de uma análise aprofundada do fenómeno aleatório em estudo. Pensamos ser este o momento ideal para dar a definição de fenómeno aleatório.

Exemplo 2

O modelo probabilístico para a soma das pintas ao lançar duas vezes um dado equilibrado (ou ao lançar dois dados) é

Soma das pintas	2	3	4	5	6	7	8	9	10	11	12
Probabilidade	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Analisemos alguns dos valores que surgem nesta tabela:

A soma é igual a 2 somente se sair 1 no primeiro lançamento (o que admitimos que tem probabilidade 1/6 de ocorrer) e sair 1 também no segundo lançamento (o que terá igualmente probabilidade 1/6 de acontecer). Tendo em conta que 1/6=0,16(6) podemos dizer que se espera que saia 1 no primeiro lançamento em 16,(6)% das vezes que se repita esta experiência e que se espera que em 16,(6)% dessas vezes saia o 1 também no segundo lançamento. Como 16,(6)% de 16,(6)% é igual a 16,(6)%×16,(6)%=2,7(7)% concluímos que deveremos esperar que a soma seja igual a 2 em 2,7(7)% das vezes que se lançar duas vezes um dado. Note-se que este 2,7(7)% não é mais do que 1/6×1/6 ou seja é o produto da probabilidade de sair o 1 no primeiro lançamento pela probabilidade de sair o 1 no segundo lançamento.

Para se ter uma soma igual a 3, ou terá de sair o 1 seguido do 2 ou o 2 seguido do 1. Espera-se que o primeiro caso ocorra em 2,7(7)% das vezes o mesmo acontecendo com o segundo caso, ou seja, espera-se que a soma seja igual a 3 em 5,(5)% das vezes. Note-se que este 5,5% pode agora ser calculado como 1/6×1/6 + 1/6×1/6 o que dá 2/36.

Facilmente se verifica que a soma que tem maior número de casos favoráveis é o 7 (1 seguido de 6 e 6 seguido de 1, 2 seguido de 5 e 5 seguido de 2, 3 seguido de 4 e 4 seguido de 3) num total de 6 casos. Uma vez que cada um dos casos tem probabilidade 1/36 de ocorrer obtém-se o valor 6/36 como probabilidade de ocorrer uma soma igual a 7 em dois lançamentos consecutivos de um dado equilibrado (ou em cada lançamento de dois dados).

Quando todos os casos são equiprováveis a probabilidade de ocorrência de um certo acontecimento pode ser calculada dividindo o número de *casos favoráveis* à ocorrência desse acontecimento pelo *total de casos possíveis*: é a chamada **Regra de Laplace**.

Exemplo 3

Tem especial interesse em estatística os modelos probabilísticos associados a situações de amostragem, isto é, a situações em que se escolhe de forma aleatória alguns indivíduos de uma certa população.

Suponhamos que numa turma com 20 alunos, 5 deles têm 15 anos, 8 têm 16 anos e 7 têm 17 anos. Não dispondo de qualquer outra informação (como, por exemplo, a forma como os alunos

estão sentados) qual o modelo probabilístico que consideraria para a idade do 1º aluno a sair da sala após o toque?
A resposta natural a esta questão é

Idade do 1º aluno a sair da sala	15	16	17
Probabilidade	5/20	8/20	7/20

Este é o modelo apropriado se admitirmos que qualquer dos alunos tem igual probabilidade de ser o primeiro a sair da sala, ou seja, que qualquer dos alunos tem probabilidade 1/20 de ser o primeiro a sair. Como há 5 alunos com 15 anos obtemos o valor 5/20 para a probabilidade de que saia primeiro um aluno de 15 anos, por exemplo.
E qual o modelo que consideraria para a idade dos 2 primeiros alunos a sair da sala?

Idade dos 2 primeiros alunos a sair da sala	(15,15)	(15,16)	(15,17)	(16,15)	(16,16)	(16,17)	(17,15)	(17,16)	(17,17)
Probabilidade	5/20×4/19	5/20×8/19	5/20×7/19	8/20×5/19	8/20×7/19	8/20×7/19	7/20×5/19	7/20×8/19	7/20×6/19

Observe-se que a saída de qualquer aluno faz alterar a composição da turma no que respeita às idades dos alunos que ainda estão na sala. Assim se o aluno que sair primeiro tiver 15 anos, dos 19 que ainda restam, teremos 4 de 15 anos, 8 de 16 anos e 7 de 17 anos. É com base neste facto que se obtêm as probabilidades indicadas.

Podemos ainda utilizar o modelo anterior para construir um modelo de probabilidade para a soma e para a média das idades dos 2 primeiros alunos a sair da sala:

Soma das idades dos 2 primeiros alunos a sair da sala	30	31	32	33	34
Média das idades dos 2 primeiros alunos a sair da sala	15	15.5	16	16.5	17
Probabilidade	5/20×4/19	2×5/20×8/19	8/20×7/19+ 2×5/20×7/19	2×8/20×7/19	7/20×6/19

Pressuposto de simetria

Qualquer um dos modelos apresentados nos três exemplos anteriores foi construído com base no chamado *pressuposto de simetria*. Este termo deriva do facto de ser devido à sua simetria física que se atribui igual probabilidade à saída de cada uma das faces de um dado. Sempre que ao realizarmos uma experiência aleatória pudermos admitir que tudo se passa como se estivéssemos a lançar um "dado" homogéneo e simétrico, então não temos razão para não atribuir igual probabilidade a todos os resultados da experiência. Analisemos o que se passa com

a experiência descrita no exemplo 3. Temos 20 alunos de diferentes idades (5 de 15 anos, 8 de 16 anos e 7 de 17 anos) e um deles, ao acaso, sai da sala. Interessa-nos atribuir uma probabilidade à idade desse aluno. Em termos probabilísticos não há qualquer diferença entre esta experiência e o lançamento de um dado homogéneo e simétrico de 20 lados, com 5 faces numeradas com o 15, 8 faces numeradas com o 16 e 7 faces numeradas com o 17. Admitindo que qualquer uma das faces (ou qualquer um dos alunos) tem igual probabilidade de sair, deveremos atribuir o valor 1/20 à probabilidade de saída de cada uma das 20 faces. Como 5 delas têm o número 15 inscrito, a probabilidade de sair uma face com o número 15 será 5/20, com o número 16 será 8/20 e com o número 17 será 7/20. O modelo obtido é pois exactamente o mesmo e deve-se salientar bem que o que esteve sempre na base do raciocínio foi o facto de se estar a atribuir igual probabilidade a cada um dos 20 resultados elementares.

Generalizando, pode então dizer-se que o modelo de probabilidade para os resultados de uma amostragem aleatória simples numa população de n indivíduos é em tudo análogo ao do lançamento de um dado homogéneo e simétrico de n faces, onde, em cada face, está representada a(s) característica(s) de interesse de cada indivíduo.

Cálculo da probabilidade de alguns acontecimentos a partir dos modelos construídos:

Exemplo 4

Qual a probabilidade de se obter uma soma superior a 8 ao lançar dois dados?

Recorrendo ao modelo construído no exemplo 2 podemos calcular facilmente esta probabilidade. Utilizando uma notação abreviada,

$$P(\text{sair soma superior a } 8) = P(\text{soma}=9) + P(\text{soma}=10) + P(\text{soma}=11) + P(\text{soma}=12)$$
$$= 10/36 \cong 0.28$$

A probabilidade da soma ser superior a 8 é, pois, aproximadamente igual a 0.28 e podemos já concluir que a probabilidade da soma ser 8 ou menos é, aproximadamente, 0.72.

Exemplo 5

Nas condições do exemplo 3 qual a probabilidade de saírem dois alunos com a mesma idade?

Que fazer quando não é possível utilizar argumentos de simetria?

Uma senhora está à espera de bebé. Será razoável atribuir igual probabilidade para que seja rapaz ou rapariga? Por outras palavras, será que esta situação é em tudo análoga à do lançamento de uma moeda equilibrada?

Na próxima jornada do campeonato de futebol o F.C.P. vai jogar com o S.B.. Será razoável dizer que tem igual probabilidade de ganhar, perder ou empatar?

Um agricultor cultiva batatas numa certa parcela de terreno. Na última colheita a produção foi de 30 arrobas. Na próxima colheita tanto poderá produzir ainda mais como não. Será então razoável dizer que tem 50% de probabilidade de vir a produzir mais do que as 30 arrobas?

Obviamente a resposta a qualquer uma das questões colocadas é negativa. Não é pelo facto de só se ter n resultados possíveis que se deve atribuir igual probabilidade a todos eles. Não sendo razoável utilizar um argumento de simetria, a solução aqui é, tal como em muitas outras ciências, recorrer à "experimentação". Teoricamente, se for possível recolher informação sobre os resultados do fenómeno aleatório que pretendemos modelar, realizando a experiência sempre nas mesmas condições, então os dados obtidos serão certamente o instrumento fundamental para se escolher um bom modelo. Acontece que quase nunca é possível realizar a experiência exactamente nas mesmas condições (o mundo está em permanente mudança...); muitas vezes não há possibilidade de realizar a experiência um número suficiente de vezes por forma a encontrar alguma regularidade na informação disponível; muitas vezes há factores desconhecidos ou não controláveis que afectam significativamente os resultados da experiência. Estes são alguns dos muitos problemas com que se debate o estatístico quando pretende sugerir um modelo de probabilidade para os resultados de um certo fenómeno aleatório. Para além disso, uma das grandes diferenças entre os modelos probabilísticos e os modelos determinísticos é que, para uma mesma situação poderá haver mais do que um modelo que interprete os dados de forma suficientemente plausível e nunca se pode dizer que "este é o modelo correcto para esta situação". Do ponto de vista do estatístico é preferível interpretar um fenómeno usando um modelo probabilístico que lhe pareça suficientemente adequado, embora não seja certamente o ideal, a cruzar os braços não fazendo nada. O que é realmente fundamental é ter consciência das limitações destes modelos, utilizando-os não como verdades absolutas mas como um meio de apoio à decisão.

Fenómenos aleatórios ***– são fenómenos cujos resultados individuais são incertos, mas para os quais se admite que se pode encontrar um padrão genérico de comportamento.***

A concordarmos com a opinião de Einstein - "Deus não joga aos dados com o Universo" - isso significaria que na realidade não existem fenómenos aleatórios mas unicamente fenómenos para os quais somos obrigados a utilizar modelos de probabilidade, por não conseguirmos conhecer exactamente as suas leis.

3.3. Modelos de probabilidade em espaços finitos. Variáveis quantitativas. Função massa de probabilidade ou distribuição de probabilidade.

Objectivos a atingir:

- ✓ Apreender as propriedades básicas de uma função massa de probabilidade.
- ✓ Identificar acontecimentos em espaços finitos.
- ✓ Saber calcular as probabilidades de alguns acontecimentos utilizando propriedades da probabilidade.

Uma vez que esta disciplina não pretende desenvolver nos alunos a capacidade de formalização, mas unicamente dotá-los de meios para mais tarde poderem interpretar e realizar alguns estudos estatísticos, deverá evitar-se dar definições que envolvam muito formalismo matemático, sendo preferível apresentar muitos exemplos para que os próprios alunos se apercebam de quais são as propriedades básicas de qualquer modelo probabilístico.

Exemplo 6

O estatístico da equipa de andebol de uma certa escola, com base no historial de jogos anteriores com o mesmo adversário, sugeriu o seguinte modelo probabilístico para o resultado final do próximo jogo:

Resultado	Vitória	Empate	Derrota
Probabilidade	0.4	0.1	0.5

O treinador, que acha que a equipa está a atravessar um bom momento de forma, é de opinião que a probabilidade de Vitória deverá ser igual a 0.6 e não 0.4. Admitindo que a probabilidade de Empate não se altera, qual é a probabilidade da equipa vir a ser derrotada?

A soma das probabilidades tem de ser igual a 1 (100%). Assim a probabilidade de derrota passará a ser igual a 0.3.

Seria possível manter a probabilidade de derrota alterando a probabilidade de empate?

Não, pois 0.6+0.5=1.1 e, para a soma de todas as probabilidades ser igual a 1, a probabilidade de empate teria de ser negativa, o que não é possível num modelo probabilístico.

Exemplo 7

O mesmo estatístico apresentou o seguinte modelo para o número de pontos marcados pela equipa

Número de pontos	De 0 a 10	De 5 a 15	Mais do que 15
Probabilidade	0.3	0.6	0.3

Será que esta tabela representa um modelo probabilístico?

A resposta aqui é mais uma vez negativa. Não pelo facto da soma das probabilidades ser superior a 1, mas sim porque os intervalos indicados para o número de pontos não são mutuamente exclusivos.

Suporte de um modelo probabilístico, Variável Aleatória e Função Massa de Probabilidade

Retomemos os quatro modelos construídos no exemplo 3:

Modelo(A)

Idade do 1º aluno a sair da sala	15	16	17
Probabilidade	5/20	8/20	7/20

Modelo(B)

Idade dos 2 primeiros alunos a sair da sala	(15,15)	(15,16)	(15,17)	(16,15)	(16,16)	(16,17)	(17,15)	(17,16)	(17,17)
Probabilidade	5/20×4/19	5/20×8/19	5/20×7/19	8/20×5/19	8/20×7/19	8/20×7/19	7/20×5/19	7/20×8/19	7/20×6/19

Modelos (C) e (D)

Soma das idades dos 2 primeiros alunos a sair da sala	30	31	32	33	34
Média das idades dos 2 primeiros alunos a sair da sala	15	15.5	16	16.5	17
Probabilidade	5/20×4/19	2×5/20×8/19	8/20×7/19+ 2×5/20×7/19	2×8/20×7/19	7/20×6/19

Os modelos (A) (C) e (D) atribuem probabilidades a grandezas quantitativas (em número finito). Ao conjunto formado por essas grandezas chamamos *suporte* do modelo. Todos eles têm, pois, suporte finito. Só mais tarde iremos ver as vantagens de se considerar modelos de suporte infinito. Mais geralmente, desde que a variável em estudo não seja de tipo quantitativo contínuo, chamaremos **suporte** de um modelo probabilístico ao conjunto formado pelos entes a que se atribuir uma probabilidade não nula. Repare-se agora que qualquer dos três modelos, (A), (C) e (D), pode ser identificado indicando o seu suporte e a correspondência entre cada ponto de suporte e a respectiva probabilidade. Ora esta correspondência é uma função real de variável real pois tem por domínio um subconjunto dos números reais e por conjunto de chegada o intervalo [0,1]. Dá-se o nome de *função massa de probabilidade* a essa correspondência, por ser uma função que atribui probabilidades a pontos de $\mathbb{R}$. Também é usual designar essa correspondência por *distribuição de probabilidade*.

Assim, para o *modelo (A)*, por exemplo, tem-se

suporte – S={15,16,17}

Função massa de probabilidade – f.m.p.=$\begin{pmatrix} 15 & 16 & 17 \\ 0.25 & 0.4 & 0.35 \end{pmatrix}$

Enquanto que, para o *modelo (D)*, se tem

suporte – S={15,15.5,16,16.5,17}

Função massa de probabilidade – f.m.p.= $\begin{pmatrix} 15 & 15.5 & 16 & 16.5 & 17 \\ \frac{20}{380} & \frac{80}{380} & \frac{126}{380} & \frac{112}{380} & \frac{42}{380} \end{pmatrix}$

Quando os modelos atribuem probabilidades a grandezas quantitativas torna-se conveniente representar essas grandezas por letras de modo a tornar a escrita mais concisa. Embora não seja regra, é usual utilizar letras maiúsculas do final do alfabeto. Também é usual chamar-lhes variáveis aleatórias.

Poderíamos então formalizar o *modelo (D)* do seguinte modo:

Seja *X* a variável aleatória que representa a média das idades dos dois alunos referidos no exemplo 3. Escrevendo

$$X=\begin{Bmatrix} 15 & 15.5 & 16 & 16.5 & 17 \\ \frac{20}{380} & \frac{80}{380} & \frac{126}{380} & \frac{112}{380} & \frac{42}{380} \end{Bmatrix}$$

indicamos claramente qual o suporte e qual a função massa de probabilidade de X. Para além disso, para identificar o acontecimento “a média das idades dos dois alunos é superior a 16” basta escrever “X>16”, tendo-se

$$P(X>16) = P(X=16.5) + P(X=17) = \frac{153}{380} \approx 0.4$$

Note-se que não se deve apresentar este tipo de caracterização se o suporte não for constituído por números reais. Tal é o caso dos modelos apresentados nos exemplos 6 e no exercício 2 . No exemplo 6 o suporte é o conjunto {Vitória, Empate, Derrota} (dizemos por isso que a variável é qualitativa) e no exercício 2 o suporte tem um elemento que não é um número mas sim um conjunto (“mais do que 5”). Como é evidente, em nenhum destes dois casos se dá o nome de função massa de probabilidade à correspondência entre os elementos do suporte e as respectivas probabilidades. No entanto, no exercício 2 faz sentido falar na variável aleatória X que representa o número de filhos de uma certa família escolhida ao acaso, pois a grandeza em causa é quantitativa. Não dispomos é de um modelo completo para essa variável.

Atentemos agora no Modelo (B): o suporte é constituído por 9 pares ordenados:

S={(15,15),(15,16),(15,17),(16,15),(16,16),(16,17),(17,15),(17,16),(17,17)}

Cada elemento do par é um número – o primeiro elemento corresponde à idade do primeiro aluno a sair da sala (X_1) e o segundo elemento à idade do segundo aluno a sair da sala (X_2). Estamos, assim, perante um par de variáveis aleatórias (X_1, X_2). Podemos fazer a sua representação de duas formas: uma análoga à utilizada para as variáveis aleatórias, isto é,

$$(X_1,X_2)=\begin{cases}(15,15) & (15,16) & (15,17) & (16,15) & (16,16) & (16,17) & (17,15) & (17,16) & (17,17)\\ \frac{20}{380} & \frac{40}{380} & \frac{35}{380} & \frac{40}{380} & \frac{56}{380} & \frac{56}{380} & \frac{35}{380} & \frac{56}{380} & \frac{42}{380}\end{cases}$$

embora seja mais usual colocar a mesma informação numa *tabela de contingência*:

X_1 \ X_2	15	16	17
15	$\frac{20}{380}$	$\frac{40}{380}$	$\frac{35}{380}$
16	$\frac{40}{380}$	$\frac{56}{380}$	$\frac{56}{380}$
17	$\frac{35}{380}$	$\frac{56}{380}$	$\frac{42}{380}$

Também neste caso se chama função massa de probabilidade à correspondência entre cada um dos pares e a respectiva probabilidade.

Exercício 1

Um aluno do 9° ano de escolaridade pode dar, no máximo, 12 faltas a Matemática. Numa certa escola fez-se um levantamento do número de faltas dadas a Matemática pelos 125 alunos do 9° ano, tendo-se obtido

N° de faltas	N° de alunos
0	1
1	6
2	15
3	12
4	20
5	25
6	12
7	5
8	5
9	7
10	5
11	10
12	?

a) Determine o número de alunos que estão tapados por faltas.

b) Construa um modelo de probabilidade para a variável X que representa o número de faltas dadas a Matemática por um dos 125 alunos do 9° ano dessa escola, escolhido ao acaso.
c) Com base no modelo da alínea anterior, calcule a probabilidade de um aluno ter menos de 3 faltas ou mais de 10.

Exercício 2

Segundo o Census 91 um possível modelo para o número de filhos em famílias monoparentais é o seguinte:

Número de filhos	0	1	2	3	4	5	Mais do que 5
Probabilidade	0.23	0.38	0.25	0.07	?	?	0.03

Sabendo que a probabilidade de uma família ter 4 filhos é um valor que está entre 0.01 e 0.02, entre que valores estará a probabilidade de uma família ter 5 filhos?

3.4. Probabilidade condicional. Árvore de probabilidades. Acontecimentos independentes.

Objectivos a atingir:

- ✓ Fazer compreender a noção de probabilidade condicional através de exemplos simples.
- ✓ Mostrar a utilidade das árvores de probabilidades como instrumento de organização de informação quando se está perante uma cadeia de experiências aleatórias.
- ✓ Ilustrar a forma de cálculo de probabilidades de acontecimentos utilizando uma árvore de probabilidades.
- ✓ Apresentar a definição de probabilidade condicional (tomando como base uma representação em diagrama de Venn de uma população classificada de forma cruzada segundo diversas categorias).
- ✓ Utilizar a definição de probabilidade condicional para formalizar a noção intuitiva de acontecimentos independentes. Apresentar a definição de acontecimentos independentes.

A noção de *probabilidade condicional* é, em geral, intuitiva para os alunos quando é aplicada no cálculo de probabilidades de cadeias de acontecimentos (ao retirar bolas de uma urna sucessivamente, sem reposição, a composição da urna altera-se e a probabilidade de se retirar certo tipo de bola depende dos tipos que saíram nas extracções anteriores). Deve-se pedir aos alunos que calculem a probabilidade de ocorrência de cadeias simples de acontecimentos aproveitando para lhes propor *esquemas em árvore* como forma de organização da informação

disponível. Deve-se então fazer ver que ao atribuírem valores às diversas probabilidades intervenientes tiveram de ter em conta quais os acontecimentos que ocorreram previamente.

Outro tipo de exemplos que conduzem facilmente à noção de probabilidade condicional são os que envolvem a "extracção" (ou escolha) ao acaso de um indivíduo de uma população cujos indivíduos estão classificados segundo os níveis de duas (ou mais) categorias (escolha ao acaso de um aluno de uma turma onde há rapazes, raparigas, filhos únicos e não filhos únicos). Representando a informação numa tabela de contingência e num *diagrama de Venn* deverá ficar claro para os alunos que a probabilidade de ocorrência de um acontecimento A tendo como informação prévia que ocorreu B (probabilidade de A condicionada pela ocorrência de B, ou probabilidade de A *condicional* à ocorrência de B) é dada pelo cociente entre a probabilidade de ocorrência dos dois acontecimentos em simultâneo e a probabilidade de ocorrência de B. Notar ainda que em situações de escolha aleatória de um indivíduo de uma população, a probabilidade de ocorrência de A condicional à ocorrência de B não é mais do que a probabilidade de ocorrência de A quando se escolhe ao acaso um indivíduo da subpopulação constituída unicamente pelos indivíduos que verificam a característica determinada pelo acontecimento B.

Considerando exemplos em que seja claro que a probabilidade de ocorrência de um acontecimento A em nada pareça poder ser afectada pela ocorrência prévia de um acontecimento B (a probabilidade de sair cara num lançamento da moeda não é afectada pelo facto de ter saído coroa no lançamento anterior) pode-se começar por dizer que esses *acontecimentos* são *independentes* e notar que se tem $P(A|B)=P(A)$. A definição de independência de dois acontecimentos deverá então ser estabelecida recorrendo a esta igualdade e à definição de probabilidade condicional. Este módulo pode terminar com a análise da independência de acontecimentos ligados à escolha aleatória de um indivíduo de uma população classificada de acordo com uma certa tabela de contingência.

Embora sem dar demasiado relevo poderão ser dados exemplos de utilização da definição de probabilidade condicional e independência em contextos genéricos onde se apresente apenas um modelo de probabilidade num certo espaço de resultados finito. Escolhendo dois acontecimentos A e B de probabilidade não nula o aluno deverá saber calcular a probabilidade de ocorrência de A condicional à ocorrência de B (e vice versa) e deverá ser capaz de verificar se A e B são ou não independentes. Também se pode utilizar a definição de probabilidade condicional na construção de novos modelos probabilísticos: considerando algum dos exemplos anteriormente apresentados (e.g., soma das pintas no lançamento de dois dados) e escolhendo um acontecimento B de probabilidade não nula (e.g., saída de soma múltipla de 3) pode-se construir um novo modelo probabilístico associado ao inicial pressupondo como informação à priori que o acontecimento B ocorreu (e.g., modelo para a soma das pintas no lançamento de dois dados sabendo que essa soma é múltipla de 3). Em situações em que faça sentido representar por X a grandeza associada ao fenómeno aleatório em estudo, é usual representar

por $X_{|B}$ essa mesma grandeza, quando se pressupõe como certa a ocorrência de B. Assim, a partir de um modelo para X e de um acontecimento B, a definição de probabilidade condicional permite construir modelos para $X_{|B}$.

Exemplo 8

Uma caixa tem 5 bolas Azuis, 8 bolas Verdes e 4 bolas Brancas. Ao retirar sucessivamente 3 bolas, qual a probabilidade da primeira ser Azul, a segunda Verde e a terceira Branca? E qual a probabilidade de saírem 3 bolas de cores diferentes?
Neste tipo de exemplos torna-se conveniente colocar um índice a indicar a ordem pela qual o acontecimento ocorreu. Assim "$\text{Azul}_{(1)}$" significa que saiu bola Azul na primeira extracção.
Usando esta notação e a regra do produto temos:

$$P(\text{Azul}_{(1)} \text{ e Verde}_{(2)} \text{ e Branca}_{(3)}) = \frac{5}{17} \times \frac{8}{16} \times \frac{4}{15}$$

P(3 bolas de cor diferente) =
= P($\text{Azul}_{(1)}$ e $\text{Verde}_{(2)}$ e $\text{Branca}_{(3)}$) + P($\text{Azul}_{(1)}$ e $\text{Branca}_{(2)}$ e $\text{Verde}_{(3)}$)
+ P($\text{Verde}_{(1)}$ e $\text{Azul}_{(2)}$ e $\text{Branca}_{(3)}$) + P($\text{Verde}_{(1)}$ e $\text{Branca}_{(2)}$ e $\text{Azul}_{(3)}$)
+ P($\text{Branca}_{(1)}$ e $\text{Azul}_{(2)}$ e $\text{Verde}_{(3)}$) + P($\text{Branca}_{(1)}$ e $\text{Verde}_{(2)}$ e $\text{Azul}_{(3)}$)
$= 6 \times \frac{5}{17} \times \frac{8}{16} \times \frac{4}{15}$

Exemplo 9

O Ricardo e a Inês estão a jogar à bisca. Neste jogo retiram-se do baralho os 8, 9 e 10 de cada naipe, restando assim 40 cartas (10 de cada naipe). No início são distribuídas 3 cartas a cada jogador. Admitindo que o Ricardo é o primeiro a receber as 3 cartas, qual é a probabilidade de lhe calhar 3 Ases? E se ele for o segundo a receber as cartas?

Facilmente se aceita que estas duas probabilidades devem ser iguais. Na realidade, tudo se passa como se tivéssemos retirado 6 cartas ao acaso do baralho e as separássemos em dois grupos de 3. A simetria de toda a experiência conduz-nos, de imediato à conclusão de que a probabilidade de estarem 3 Ases em qualquer destes dois grupos é igual. Vamos, no entanto, verificar esse facto admitindo que as cartas vão sendo distribuídas uma a uma, em sequência e utilizando a regra do produto.

Sendo o Ricardo o primeiro a receber as cartas, a probabilidade de que a primeira seja um Ás é, obviamente, 4/40, pois há 4 Ases no total das 40 cartas. Tendo recebido um Ás na primeira carta,

a probabilidade de que a segunda também seja um Ás passa a ser 3/39, pois já só há 3 Ases num total de 39 cartas. Finalmente, tendo já dois Ases na mão, a probabilidade de vir a receber um terceiro Ás é 2/38. Tem-se, assim

$$\textbf{P(3 Ases quando é o primeiro a receber as cartas) = } \frac{4}{40}\times\frac{3}{39}\times\frac{2}{38}=\frac{1}{2470}$$

Se o Ricardo for o segundo a receber as cartas a forma de cálculo altera-se pois tudo depende das 3 cartas que a Inês tiver recebido. Assim, se ela tiver recebido 2 ou 3 Ases o Ricardo já não poderá receber os 3 Ases. Por outro lado, se a Inês não tiver recebido nenhum Ás, então a probabilidade do Ricardo vir a receber 3 Ases é $\frac{4}{37}\times\frac{3}{36}\times\frac{2}{35}$, enquanto que, se a Inês tiver recebido um Ás, a probabilidade do Ricardo vir a receber 3 Ases é $\frac{3}{37}\times\frac{2}{36}\times\frac{1}{35}$. Para calcular a probabilidade do Ricardo receber 3 Ases, sendo o segundo a receber as cartas, temos de somar as probabilidades das 2 sequências de acontecimentos que são favoráveis a que tal aconteça: "Inês não recebe Ases" seguido de "Ricardo recebe 3 Ases"; "Inês recebe 1 Ás" seguido de "Ricardo recebe 3 Ases".

Sempre que temos uma sequência de acontecimentos a sua probabilidade obtém-se multiplicando sucessivamente as respectivas probabilidades (regra do produto) tendo sempre em conta de que modo a ocorrência de cada um afecta a probabilidade de ocorrência dos seguintes. Assim, para a primeira sequência de acontecimentos tem-se

P("Inês não recebe Ases" seguido de "Ricardo recebe 3 Ases") =

$$= \frac{36}{40}\times\frac{35}{39}\times\frac{34}{38}\times\frac{4}{37}\times\frac{3}{36}\times\frac{2}{35}=\frac{34\times4\times3\times2}{40\times39\times38\times37}.$$

Antes de calcularmos a probabilidade associada à segunda sequência, temos de calcular a probabilidade da Inês ter um Ás. Uma das formas de calcular esta probabilidade (evitando o recurso ao cálculo combinatório) é pensando que ela ou recebe esse Ás na primeira carta, ou na segunda ou na terceira. Com base neste tipo de esquema de raciocínio, obtém-se

P("Inês receber 1 Ás") = $\frac{4}{40}\times\frac{36}{39}\times\frac{35}{38}+\frac{36}{40}\times\frac{4}{39}\times\frac{35}{38}+\frac{36}{40}\times\frac{35}{39}\times\frac{4}{38}=3\times\frac{4}{40}\times\frac{36}{39}\times\frac{35}{38}$

Utilizando este resultado temos, então

P("Inês recebe 1 Ás" seguido de "Ricardo recebe 3 Ases") =

$$=3\times\frac{4}{40}\times\frac{36}{39}\times\frac{35}{38}\times\frac{3}{37}\times\frac{2}{36}\times\frac{1}{35}=\frac{3\times4\times3\times2}{40\times39\times38\times37}$$

Somando as probabilidades obtidas para as duas sequências de acontecimentos obtemos finalmente a probabilidade do Ricardo receber 3 Ases quando é o segundo a receber as cartas

P(3 Ases quando é o segundo a receber as cartas) =

$$= \frac{37\times 4\times 3\times 2}{40\times 39\times 38\times 37} + \frac{3\times 4\times 3\times 2}{40\times 39\times 38\times 37} = \frac{4\times 3\times 2}{40\times 39\times 38} = \frac{1}{2470}$$

o que confirma o que foi dito inicialmente, isto é, a probabilidade não se altera com a ordem por que são distribuídas as cartas.

Árvore de Probabilidades

Exemplo 10

O Luís mora longe da escola e por isso chega muitas vezes atrasado à primeira aula. Na realidade ele levanta-se praticamente sempre a horas (só em 5% dos dias é que volta a adormecer depois do despertador tocar), mas como tem de apanhar um autocarro e depois um comboio tem muitas vezes problemas se o autocarro se atrasar e não conseguir apanhar o comboio que lhe permite chegar a horas. O Luís resolveu tomar nota do que ia acontecendo ao longo de vários dias e chegou aos seguintes resultados:

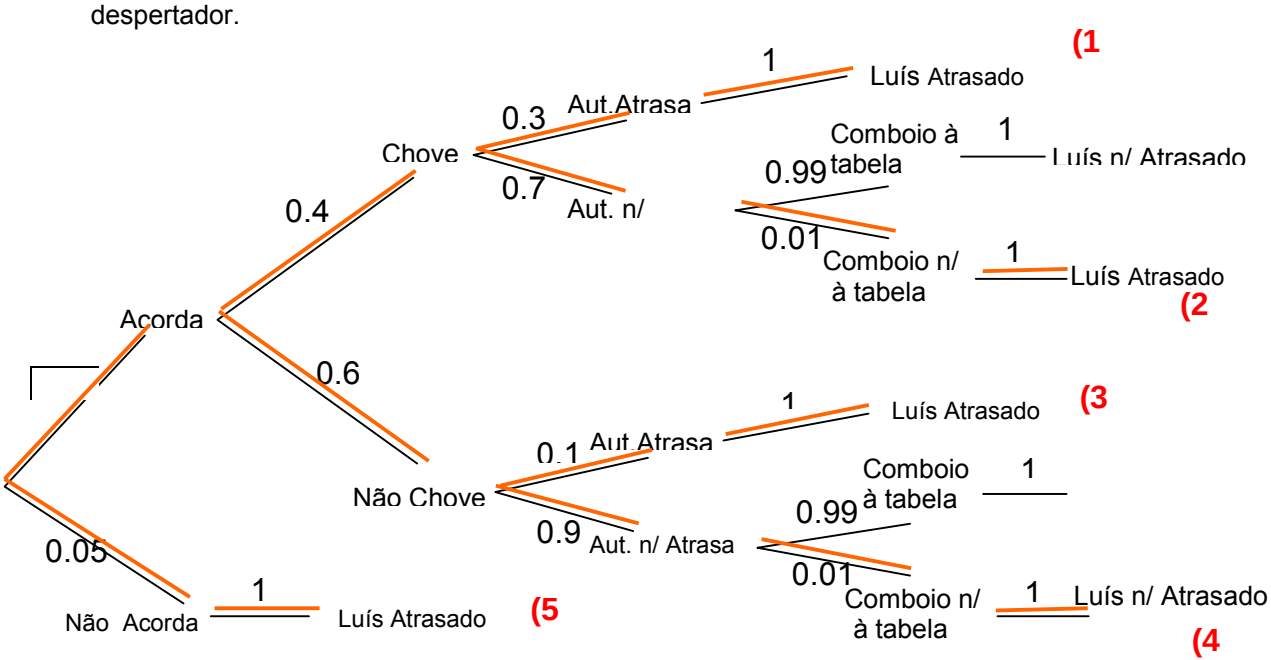
Se estiver a chover (o que acontece em 40% dos dias), o autocarro atrasa-se com uma probabilidade de 0.30. Caso contrário, essa probabilidade baixa para 0.1.

Quase nunca há problemas com o trajecto de comboio mas, mesmo assim, em 1% dos dias ele não consegue chegar à tabela e isso é o suficiente para que o Luís chegue atrasado.

É claro que, se não se levantar assim que o despertador toca, então não há nada a fazer e chega mesmo atrasado.

Num ano lectivo com 180 dias de aulas, em quantas se espera que o Luís chegue atrasado?

Para dar resposta a esta questão vamos começar por representar num esquema em árvore a cadeia de acontecimentos que condicionam a chegada do Luís à escola desde que toca o despertador.



São 5 os trajectos que conduzem a “Luís Atrasado”. Para calcular a probabilidade (total) de ocorrência deste acontecimento basta calcular a probabilidade associada a cada trajecto que a ele conduz, partindo do nó da árvore, e somar todas essas probabilidades. Para calcular a probabilidade de cada trajecto basta ir multiplicando, sucessivamente, as probabilidades que surgem em cada uma das passagens. Assim, a probabilidade associada ao trajecto (1) é

$$0.95\times0.4\times0.3\times1 = 0.114$$

a probabilidade associada ao trajecto (2) é

$$0.95\times0.4\times0.7\times0.01\times1 = 0.00266$$

Luís Atrasado

a probabilidade associada ao trajecto (3) é

$$0.95\times0.6\times0.1\times1 = 0.057$$

a probabilidade associada ao trajecto (4) é

$$0.95\times0.6\times0.9\times0.01\times1 = 0.0051$$

e a probabilidade associada ao trajecto (5) é

$$0.05\times1 = 0.05$$

Somando todos estes valores obtemos finalmente a probabilidade do Luís chegar atrasado à primeira aula

$$P(\text{Luís Atrasado}) = 0.114+0.00266+0.057+0.0051+0.05 = 0.22876$$

Multiplicando este valor pelos 180 dias de aulas obtém-se o valor 41.1768, isto é, com base nas probabilidades atribuídas a cada um dos acontecimentos intervenientes, e admitindo que mais nada pode causar o atraso do Luís, espera-se que ele chegue atrasado em cerca de 41 dos 180 dias de aulas.

Definição de Probabilidade Condicional

Nos exemplos anteriores utilizou-se diversas vezes a noção de probabilidade condicional: o valor que atribuímos à probabilidade de alguns dos acontecimentos esteve dependente da ocorrência ou não de outros. No exemplo do jogo da bisca vimos que a probabilidade do Ricardo receber 3 Ases quando a Inês não recebeu nenhum Ás é diferente da probabilidade de receber os 3 Ases quando a Inês já recebeu 1. Representemos por I o número de ases recebidos pela Inês e por R o número de Ases recebidos pelo Ricardo. Podemos então escrever

$P(R=3 \textit{ sabendo que } I=0) = \frac{4}{37}\times\frac{3}{36}\times\frac{2}{35}$ $P(R=3 \textit{ sabendo que } I=1) = \frac{3}{37}\times\frac{2}{36}\times\frac{1}{35}$

P(R=3 *sabendo que* I=2) = 0 P(R=3 *sabendo que* I=3) = 0

Quando temos dois acontecimentos A e B e queremos representar a probabilidade de "A ocorrer *sabendo que* B ocorreu" (ou "A ocorrer *dado que* B ocorreu" ou "A ocorrer *após* a ocorrência de B") utilizamos uma barra vertical, escrevendo simplesmente P(A|B) e lemos *Probabilidade de A se B* ou *Probabilidade de A condicional a B*.

Em todos os exemplos apresentados até ao momento, por estarmos perante cadeias ou sequências de acontecimentos, as diversas probabilidades condicionais ou eram dadas à partida (no exemplo do atraso do Luís, é dito no início que P(autocarro atrasar | está a chover) = 0.3) ou eram facilmente calculadas porque as ocorrências anteriores se tinham limitado a alterar a composição da "caixa de onde estávamos a tirar as bolas" (exemplo 8) :

$P(Azul_{(1)} \text{ e } Verde_{(2)} \text{ e } Branca_{(3)}) = P(Azul_{(1)})\, P(Verde_{(2)} | Azul_{(1)})\, P(Branca_{(3)} | Azul_{(1)} \text{ e } Verde_{(2)})$

Vamos agora ver outro tipo de exemplos onde também faz sentido falar em probabilidade condicional e que, para além disso nos permite vir a dar para ela uma definição.

Exemplo 11

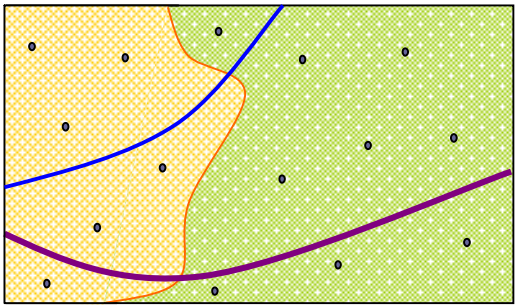
Perguntou-se aos 15 alunos de uma turma (9 rapazes e 6 raparigas) de qual dos dois tipos de música gostavam mais: "Metálica" ou "Rap". Também tinham em alternativa dizer que não gostavam de nenhum destes dois tipos de música. Como resultado deste inquérito obteve-se a seguinte tabela:

	Metálica	Rap	Nenhuma
Masculino	3	5	1
Feminino	1	2	3

Escolhendo um destes alunos ao acaso, a probabilidade de ele gostar mais de música Rap é, obviamente, 7/15. No entanto, se se disser que esse aluno escolhido é do sexo feminino, essa probabilidade altera-se para 2/6. Esta última probabilidade é condicional pois é-nos dada uma informação a priori que reduz a população inicial. Temos então, P(Rap)=7/15 e P(Rap | sexo Feminino)=2/6. Note-se agora que este mesmo valor se obtém se dividirmos a probabilidade do aluno escolhido verificar, simultaneamente, ser de sexo Feminino e gostar mais de música Rap, que é 2/15, pela probabilidade de ele ser de sexo feminino, que é 6/15. Podemos então dizer que

$$P(\text{Rap} \mid \text{sexo Feminino}) = \frac{P(\text{sexo Feminino e Rap})}{P(\text{sexo Feminino})}$$

Esta mesma relação é bem visualisável representando os diversos conjuntos de interesse num diagrama de Venn



A zona a laranja contém 6 pontos que representam as 6 raparigas da turma, a zona a verde contém 9 pontos que representam outros tantos rapazes e as linhas azul e grená separam o conjunto dos 15 alunos pelo tipo de música que preferem (abaixo da linha grená estão os que preferem música Metálica; entre essa linha e a azul estão os que preferem música Rap; acima da linha azul estão os que não preferem qualquer destes dois tipos de música). Facilmente se compreende que a probabilidade de escolher um "ponto" da região correspondente à música Rap é distinta consoante se considere como zona de escolha todo o universo, a zona laranja ou a zona verde. A primeira vai corresponder à probabilidade total e as outras duas a probabilidades condicionais.

Note-se ainda que a probabilidade de ocorrência simultânea de dois dos acontecimentos se obtém dividindo o número de pontos que estão na intersecção dos respectivos conjuntos pelo número total de pontos.

Temos então a seguinte definição de ***probabilidade condicional***:

Sendo B um acontecimento possível

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Nos problemas em que a probabilidade condicional é conhecida à partida, esta igualdade permite-nos calcular a probabilidade da ocorrência simultânea dos dois acontecimentos, pois tem-se

$$P(A \cap B) = P(B) \times P(A \mid B) \qquad (1)$$

Aliás, é esta a igualdade que temos utilizado sistematicamente quando aplicámos a regra do produto para sequências de acontecimentos.

Acontecimentos Independentes

Retomemos os dados do exemplo 8, supondo agora que, de cada vez que se retira uma bola, se regista a sua cor e se volta a colocar a bola na caixa (extracção com reposição). Qual será agora a probabilidade de retirar primeiro uma bola Azul, depois uma Verde e a seguir uma Branca?

Pela regra do produto

$$P(\text{Azul}_{(1)} \text{ e Verde}_{(2)} \text{ e Branca}_{(3)}) = P(\text{Azul}_{(1)})\, P(\text{Verde}_{(2)} \mid \text{Azul}_{(1)})\, P(\text{Branca}_{(3)} \mid \text{Azul}_{(1)} \text{ e Verde}_{(2)})$$

Acontece que, neste caso, seja qual for a cor das bolas extraídas anteriormente e seja qual for a ordem de extracção, a probabilidade associada a cada uma das cores é sempre a mesma: 5/17 para a cor Azul, 8/17 para a Verde e 4/17 para a Branca. Assim sendo,

$P(Verde_{(2)}| Azul_{(1)}) = P(Verde) = 8/17$

e

$P(Branca_{(3)}| Azul_{(1)} \text{ e } Verde_{(2)}) = P(Branca) = 4/17$

Mais geralmente, sempre que se tiver P(A|B)=P(A), diremos que os acontecimentos A e B são independentes. A definição formal de independência é, no entanto, a que se obtém substituindo na igualdade (1), P(A|B) por P(A). Mais precisamente dois acontecimentos A e B são independentes se e só se

$$P(A \cap B) = P(A) \times P(B)$$

Note-se que tem grande importância em estatística a amostragem aleatória com reposição (ou independente) principalmente porque o facto de neste caso as probabilidades não serem afectadas pelas ocorrências anteriores vem facilitar imenso os cálculos e permitir que se estabeleçam resultados de extrema utilidade em previsão, como são, por exemplo alguns intervalos de confiança e testes de hipóteses.

Exercício 3

Um jogo de computador tem três níveis cada vez mais difíceis de passar. O Hugo já se considera bastante treinado nesse jogo e por isso considera que a sua probabilidade de passar o nível 1 é de 95%; quando consegue passar o nível 1, a probabilidade de também conseguir passar o nível 2 é de 80% e, ultrapassados os dois primeiros níveis, a probabilidade de conseguir chegar ao fim do nível 3 é de 70%.

O Hugo vai iniciar um novo jogo. Qual é a probabilidade que ele tem de o conseguir terminar? E qual é a probabilidade de chegar ao fim do nível 2?

3.5. Probabilidade total. Regra de Bayes.

Objectivos a atingir:

- ✓ Introduzir os alunos nas técnicas Bayesianas, que se baseiam no seguinte princípio: começa-se por atribuir uma probabilidade a um acontecimento, tendo em consideração a informação disponível – probabilidade à priori, e posteriormente, mediante nova informação entretanto adquirida, obtém-se uma nova probabilidade para esse acontecimento – probabilidade à posteriori, que se pode entender como uma correcção da probabilidade anteriormente dada.

Recorrendo novamente a um exemplo de escolha aleatória de um indivíduo duma população mostrar que se pode calcular a probabilidade desse indivíduo verificar a característica A desde que se conheça a composição percentual da população no que respeita a uma sua subdivisão em subpopulações disjuntas e se conheçam as probabilidades de ocorrência de A quando se restringe a população inicial a cada uma dessas subpopulações. Ilustrar com um diagrama de Venn e apresentar a fórmula de cálculo da probabilidade de A como sendo uma soma das probabilidades de A para cada subpopulação, ponderada pelo coeficiente dessa subpopulação na composição percentual da população total. Esta propriedade é conhecida como da *Probabilidade total*. Dar exemplos também em situações de cadeia de acontecimentos e em situações típicas de causa / efeito. Neste último caso utilizar o termo "probabilidade *à priori*" que deverá motivar o aparecimento da regra de Bayes como uma forma de calcular uma "probabilidade *à posteriori*". Conhecendo à partida as probabilidades de um certo efeito A ser originado por cada uma das n causas possíveis e mutuamente exclusivas e conhecendo o modelo de probabilidade (modelo das probabilidades *à priori)* para essas causas, a *regra de Bayes* permite calcular a "probabilidade *à posteriori*" – após a ocorrência de A – de ter sido uma determinada, a causa que originou A. Devem ser dados muitos exemplos neste contexto e posteriormente deve ser alargada a aplicação da regra de Bayes a situações em que não faça muito sentido dizer que se esteja perante relações de causa/efeito.

Probabilidade total

Este é um termo já utilizado em alguns dos exemplos dados anteriormente mas de que só agora se irá apresentar a definição. Desses exemplos o mais sugestivo é talvez o do "atraso do Luís". Como se viu, para calcular a probabilidade do Luís chegar atrasado à primeira aula somámos as probabilidades correspondentes a cada uma das 5 sequências de acontecimentos assinaladas na árvore e que terminavam em "Luís Atrasado". Vamos chamar S1 à sequência dos acontecimentos que, na árvore, antecedem o primeiro "Luís Atrasado", ou seja, S1=(Acorda e Chove e Aut. Atrasa). De igual modo S2, S3, S4 e S5 constituem as restantes sequências de acontecimentos, possíveis causadoras do atraso do Luís (S5, por exemplo, é, simplesmente, "Não Acorda"). Note-se que todas elas são mutuamente exclusivas e que se está a admitir que mais nada poderá causar o atraso do Luís. Analisando o modo como calculámos a probabilidade de "Luís Atrasado" verificamos que

P(Luís Atrasado)=P(S1 e Luís Atrasado)+P(S2 e Luís Atrasado)+...+P(S5 e Luís Atrasado)

Uma **probabilidade total** é sempre calculada como uma soma de probabilidades parciais correspondentes à intersecção do acontecimento de interesse com outros mutuamente exclusivos e exaustivos (utilizando a linguagem da teoria de conjuntos, esses acontecimentos têm de formar uma partição do universo).

Exemplo 12

Uma fábrica tem 3 máquinas de fazer parafusos. A máquina A é muito antiga e por isso produz poucos parafusos (apenas 20% da produção global) e a sua taxa de parafusos defeituosos é de 5%; a máquina B é a mais rápida, produzindo 50% da produção global mas, em contrapartida tem uma taxa de defeituosos igual a 3% que é o dobro da taxa de parafusos defeituosos entre os produzidos pela máquina C. Qual é a percentagem de parafusos defeituosos na produção global da fábrica?

Para calcular a probabilidade total de, escolhendo um parafuso ao acaso, ele ser defeituoso, temos de somar as probabilidades parciais correspondentes à ocorrência simultânea de ser defeituoso e proveniente de cada uma das três máquinas

$$P(\text{Def.}) = P(\text{A e Def.}) + P(\text{B e Def.}) + P(\text{C e Def.})$$

Mas, por (1), sabemos que P(Def. e A)=P(A) P(Def.| A), P(Def. e B)=P(B) P(Def.| B) e P(Def. e C)= P(C) P(Def.| C). Logo

$$P(\text{Def.}) = P(A)\,P(\text{Def.}|\,A) + P(B)\,P(\text{Def.}|\,B) + P(C)\,P(\text{Def.}|\,C)$$

$$= 0.2\times0.04 + 0.5\times0.03 + 0.3\times0.015 = 0.0275$$

Concluímos assim que 2.75% dos parafusos produzidos por esta fábrica são defeituosos.

Regra de Bayes

A regra de Bayes obtém-se substituindo na expressão que define a probabilidade condicional, o numerador do segundo membro por uma expressão equivalente (que é a que se obtém de (1) trocando A por B), vindo

$$P(A\,|\,B) = \frac{P(B\,|\,A)\;P(A)}{P(B)}$$

Esta expressão é utilizada para calcular, *a posteriori*, qual a probabilidade de ter sido a causa A a dar origem ao efeito B. Note-se que P(B | A) representa a probabilidade, *a priori*, de A dar origem ao efeito B.

Voltemos novamente ao exemplo 9. Já vimos que o atraso do Luís na primeira aula pode ser devido, unicamente, a 5 causas mutuamente exclusivas (S1, S2, S3, S4 e S5). Num certo dia o Luís chegou atrasado à primeira aula. Qual é a probabilidade de ele se ter deixado adormecer? Vamos então calcular a probabilidade, *a posteriori*, de ter sido S5 a causa do atraso do Luís, isto é, P(S5 | Luís Atrasado). Utilizando a regra de Bayes, vem

$$P(\text{S5}\,|\,\text{Luís Atrasado}) = \frac{P(\text{Luís Atrasado}\,|\,\text{S5})\;P(\text{S5})}{P(\text{Luís Atrasado})}$$

Recordando que S5 é o acontecimento “Não Acorda”, cuja probabilidade é 0.05, e que quando o Luís não acorda então chega atrasado com probabilidade 1, verificamos que dispomos de todos os elementos para substituir no numerador. Quanto ao denominador, ele não é mais do que a probabilidade total do Luís chegar atraso – 0.22876. Logo

$$P(\text{S5}\,|\,\text{Luís Atrasado}) = \frac{1\times0.05}{0.22876} \approx 0.218$$

3.6. Valor médio e variância populacional

Objectivos a atingir:

- ✓ Fazer a distinção entre valor médio (ou média) populacional e média amostral e também, de modo idêntico, para a variância e outras características já referidas no estudo descritivo de amostras.
- ✓ Alargar a noção de população como um conceito subjacente a um modelo de probabilidade.
- ✓ Apresentar, de forma justificada, a fórmula de cálculo do valor médio para modelos quantitativos de espaço de resultados finito.

Este é o módulo fundamental para a compreensão dos tópicos que irão ser tratados no capítulo da inferência estatística. Mais precisamente, naqueles capítulos irão ser dados resultados que irão permitir fazer certas afirmações (probabilísticas) sobre características de interesse numa população, tendo como base unicamente a informação constante numa pequena parte dessa população (amostra). É sabido que uma das mais importantes características (ou parâmetro) de interesse numa população cuja variável em estudo seja quantitativa, é a sua média. Esta, para não ser aqui confundida com a média de uma amostra que se retire dessa população, será designada por *valor médio* e terá uma representação convencional através da letra μ. A variância (desvio padrão) correspondente a todos os valores da variável em estudo na população designa-se por *variância* (desvio padrão) populacional e representa-se por σ^2 (σ). Qualquer das outras características amostrais, apresentadas no capítulo da Estatística (mediana, quantis, etc.) quando referidas a todos os elementos de uma população, serão designadas pelo mesmo nome a que se acrescenta o termo "populacional" (mediana populacional, quantis populacionais, etc.). A própria probabilidade de um indivíduo da população verificar certa propriedade pode ser encarada como um parâmetro de interesse que terá como contrapartida amostral a frequência com que essa propriedade foi observada (na amostra).

Com o objectivo de alargar o conceito de *população* como algo de subjacente a qualquer modelo de probabilidade comece-se por considerar uma população finita (por exemplo, os alunos da turma) e escolha-se uma sua característica quantitativa de interesse (por exemplo, o número de irmãos). Estabelecendo o modelo de probabilidade apropriado para o valor observado dessa característica quando se escolhe um indivíduo ao acaso dessa população (modelo de probabilidade para o número de irmãos de um aluno escolhido ao acaso na turma) deve-se então provar que o valor médio da população escolhida (média do número de irmãos de todos os alunos da turma) também pode ser calculado a partir do modelo, multiplicando cada valor do espaço de resultados pela respectiva probabilidade e somando todos os valores obtidos. O passo seguinte é apresentar a definição de valor médio de um modelo de probabilidade com suporte

finito (mesmo que esse modelo não se refira ao valor observado de uma característica quantitativa de um indivíduo escolhido ao acaso numa certa população). Faz então sentido falar em valor médio de qualquer variável quantitativa a que se tenha associado um modelo de probabilidade (a soma média, ou valor esperado da soma das pintas ao lançar duas vezes um dado não é mais do que o valor médio do modelo de probabilidade respectivo). Embora sem apresentar as fórmulas de cálculo deve-se dizer que também se pode definir variância, mediana e quantis de um modelo de probabilidade.

Neste momento deve ficar claro para os alunos que se utilizam termos análogos em três contextos distintos: amostra, população, modelo de probabilidade. Seria agora de extrema importância que se conseguisse fazer compreender de que modo é possível alargar o conceito de população de modo a que seja indiferente utilizar, por exemplo, o termo valor médio para população ou valor médio para o modelo. Ao lançar duas vezes um dado não faz, à partida, muito sentido, dizer que a soma das pintas é um valor observado de uma certa característica de um indivíduo escolhido ao acaso de uma população. Tal só é possível de forma abstracta admitindo que a população é constituída por todos os resultados ocorridos em todos os duplos lançamentos que se imagine que foram efectuados com o dado. A população associada a um modelo probabilístico deixa de ser constituída por indivíduos para passar a ser uma colecção de números (eventualmente infinita) que tem a particularidade de ter uma composição percentual respeitante a cada um dos diversos valores que a compõe, determinada pelo referido modelo. Com base nesta população (conceptual) o modelo probabilístico refere-se sempre ao valor observado de uma extracção ao acaso de um elemento dessa população.

Exemplo 13

A média das idades dos 20 alunos mencionados no exemplo 3 é (uma vez que 5 deles têm 15 anos, 8 têm 16 e 7 têm 17 anos)

$$\frac{5\times 15+8\times 16+7\times 17}{20}=16.1$$

Por outro lado, o modelo para a variável aleatória X que representa a idade de um aluno escolhido ao acaso entre os 20 é

$$X=\begin{cases} 15 & 16 & 17 \\ \frac{5}{20} & \frac{8}{20} & \frac{7}{20} \end{cases}$$

O valor médio deste modelo calcula-se multiplicando cada um dos elementos do suporte pela respectiva probabilidade e somando os valores assim obtidos. O valor médio de uma variável aleatória representa-se por E(X) ou por μ_X . Neste caso, o valor médio do modelo coincide com a média da população de onde se escolheu ao acaso um elemento, pois,

$$\mu_X = 15\times\frac{5}{20}+16\times\frac{8}{20}+17\times\frac{7}{20}=16.1$$

3.7. Espaços de resultados infinitos. Modelos discretos e modelos contínuos.

Objectivos a atingir:

- ✓ Mostrar o interesse em adoptar modelos com suporte não finito em situações onde o conjunto de resultados possíveis não seja conhecido na sua totalidade ou seja demasiado extenso.
- ✓ Calcular probabilidades de acontecimentos a partir de alguns modelos contínuos simples.

Através da discussão de alguns exemplos comuns (nº de filhos das famílias portuguesas, alturas de todos os rapazes da escola, tempo de duração de um equipamento, etc.) alertar para as vantagens de se escolher um modelo de suporte infinito.

Seguidamente deve-se referir o facto de que qualquer função cujo gráfico nunca passe abaixo do eixo dos xx, e tal que a área compreendida entre o gráfico da função e o eixo dos xx seja igual a uma unidade, permite construir um modelo de probabilidade no conjunto dos números reais. Na realidade, para calcular a probabilidade de qualquer intervalo basta calcular a área determinada por esse intervalo, entre o eixo dos xx e a curva. Será interessante dar exemplos sugestivos com funções de densidade de probabilidade cujos gráficos tenham áreas simples de calcular. Pode-se mencionar o modelo uniforme e o modelo exponencial.

Para estudar modelos discretos de suporte infinito é necessário que os alunos consigam calcular a soma de algumas séries. A sugestão que poderemos fazer neste campo é a de que se apresentem alguns modelos simples, nomeadamente o modelo Geométrico e o modelo Poisson, referindo algumas situações em que são utilizados.

Qual o interesse em utilizar modelos de probabilidade com suporte infinito?

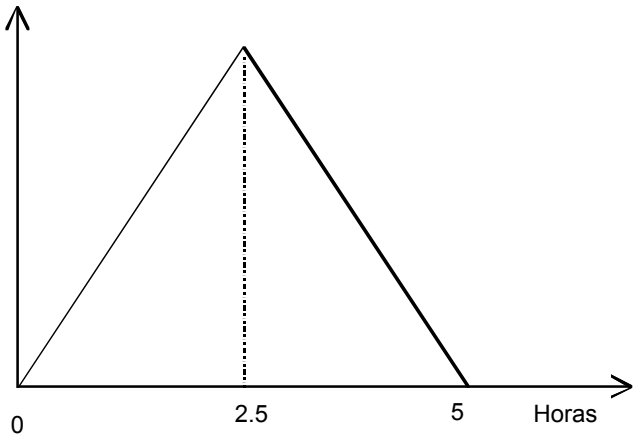
Quando anteriormente referimos o exemplo do número de filhos de famílias monoparentais dissemos que não se tratava de um modelo de probabilidade porque no suporte estava o “ente” “Mais do que 5” o qual não é um número mas sim um conjunto. A dificuldade aqui é que para se adoptar um modelo de suporte finito teríamos de saber qual o número máximo de filhos que uma família pode ter ou, pelo menos, escolher um número suficientemente elevado de modo a que fosse nula a probabilidade de uma família ter mais do que esse número de filhos. Ficaríamos com um modelo de suporte tão extenso que acabaria por ser de utilização mais difícil que alguns dos modelos de suporte infinito de que iremos falar de seguida. Como este há muitos exemplos de situações em que são claras as vantagens em utilizar os chamados modelos discretos de suporte infinito (número de carros que passam numa portagem por hora, número de clientes que entram numa loja por dia, número de partículas poluentes por unidade de volume num rio, etc.). Os modelos discretos de suporte infinito mais utilizados são o de Poisson e o Geométrico.

No entanto, é na modelação de variáveis de tipo contínuo que são claramente óbvias as vantagens dos modelos cujo suporte seja, ou um intervalo, ou todo o conjunto dos números reais.

Quando os dados de que dispomos se referem a alturas, pesos, tempos de vida, ou outras grandezas que podem assumir qualquer valor dum certo intervalo então o modelo de probabilidade a adoptar deverá ser descrito por uma função real de variável real que tenha por domínio o intervalo (ou os intervalos) onde consideramos que a probabilidade deverá ser não nula. Esta função chama-se *função de densidade* e deverá ter um gráfico que nunca passe abaixo do eixo dos xx. A área total entre o gráfico e o eixo dos xx deverá ser igual à unidade, sendo a probabilidade da variável assumir valores num certo intervalo [a,b] dada pela área determinada pelo gráfico da função de densidade e esse intervalo. Aliás, a razão de ser do nome "função de densidade" vem exactamente do facto de ela traduzir quais são as zonas de maior ou menor densidade de observação quando se recolherem amostras provenientes de uma população que seja bem descrita por esse modelo.

Exemplo 14

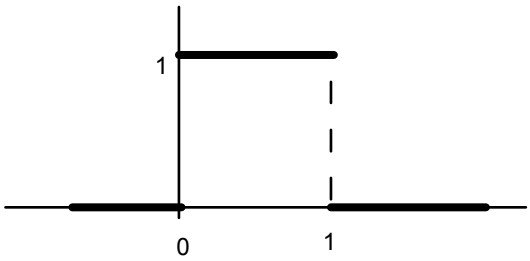
Como todos os alunos, o André estuda mais na véspera dos pontos do que nos restantes dias. No entanto, nunca estuda mais do 5 horas e com maior frequência estuda entre 2 e 3 horas. Com base nesse facto sugeriu o seguinte modelo para o número de horas de estudo na véspera dos pontos:



De notar que a informação apresentada é suficiente para identificar a função de densidade. Começando por determinar a altura do triângulo de modo a que a sua área seja unitária, ficamos com todos os elementos para determinar a equação das duas rectas que definem a função. É agora só uma questão de cálculo de áreas a determinação de probabilidades associadas a esta situação (probabilidade de estudar entre 2 e 3 horas, probabilidade de estudar mais de 4 horas, etc.).

Modelo Uniforme e Modelo Exponencial

O modelo Uniforme é o que atribui igual densidade a qualquer zona de um certo intervalo. Está por isso na base dos geradores de números (pseudo) aleatórios (NPA's) e na simulação de amostras provenientes de um certo modelo.



(Gráfico da Função de densidade do modelo Uniforme no intervalo [0,1])

Os algoritmos de geração de números pseudo-aleatórios estão concebidos de modo a que ao considerar uma qualquer sequência de números gerados se obtenha aproximadamente a mesma proporção de observações em subintervalos de igual amplitude do intervalo [0,1]. Assim, por exemplo, se se fizer correr o algoritmo 100 vezes, é de esperar que caiam 25 dos números gerados em cada quarto do intervalo [0,1]. Na tabela seguinte está listada uma sequência de 100 NPA's obtida através do gerador RAND do software Excel.

0,842050	0,406320	0,848744	0,810469	0,789583
0,965131	0,676239	0,722927	0,825587	0,702971
0,761648	0,552387	0,079614	0,298300	0,087455
0,359825	0,208420	0,098150	0,818893	0,103532
0,054705	0,102768	0,147229	0,557920	0,996667
0,466613	0,493374	0,150888	0,540352	0,480287
0,814300	0,638416	0,086141	0,007840	0,109918
0,449515	0,090759	0,197460	0,209145	0,713230
0,901502	0,552418	0,466389	0,221584	0,623757
0,862762	0,507097	0,613583	0,389183	0,129629
0,395195	0,415666	0,210044	0,379011	0,302539
0,420519	0,469764	0,053714	0,478208	0,444822
0,124664	0,765629	0,737348	0,696311	0,806147
0,537707	0,451921	0,702749	0,683382	0,377823
0,033277	0,523063	0,908485	0,708764	0,196290
0,024371	0,213326	0,442821	0,983754	0,970551
0,558313	0,283191	0,153907	0,655705	0,995760
0,087859	0,429387	0,735276	0,890680	0,569285
0,069915	0,221549	0,358037	0,578713	0,161851
0,774156	0,039495	0,490216	0,755072	0,753139

Como se pode verificar por contagem, esta lista inclui 30 números no intervalo [0,0.25], 24 números nos intervalos]0.25,0.5] e]0.5,0.75] e 22 números no intervalo]0.75,1]. Embora haja métodos estatísticos para avaliar se são ou não significativas as diferenças entre estas frequências observadas e as frequências esperadas (25 – 25 – 25 – 25), facilmente a nossa sensibilidade aceita que estes resultados não contradizem o que se esperaria de uma escolha ao acaso de 100 números do intervalo [0,1].

Esta lista de NPA's conduz de imediato à seguinte simulação de 100 lançamentos de uma moeda equilibrada (NPA$\leq$0.5 $\rightarrow$ Cara; NPA>0.5 $\rightarrow$ Coroa):

Coroa	Cara	Coroa	Coroa	Coroa
Coroa	Coroa	Coroa	Coroa	Coroa
Coroa	Coroa	Cara	Cara	Cara
Cara	Cara	Cara	Coroa	Cara
Cara	Cara	Cara	Coroa	Coroa
Cara	Cara	Cara	Coroa	Cara
Coroa	Coroa	Cara	Cara	Cara
Cara	Cara	Cara	Cara	Coroa
Coroa	Coroa	Cara	Cara	Coroa
Coroa	Coroa	Coroa	Cara	Cara
Cara	Cara	Cara	Cara	Cara
Cara	Cara	Cara	Cara	Cara
Cara	Coroa	Coroa	Coroa	Coroa
Coroa	Cara	Coroa	Coroa	Cara
Cara	Coroa	Coroa	Coroa	Cara
Cara	Cara	Cara	Coroa	Coroa
Coroa	Cara	Cara	Coroa	Coroa
Cara	Cara	Coroa	Coroa	Coroa
Cara	Cara	Cara	Coroa	Cara
Coroa	Cara	Cara	Coroa	Coroa

Ou à seguinte simulação da soma obtida em 100 lançamentos de dois dados equilibrados.

Soma	Soma	Soma	Soma	Soma
10	6	10	9	9
11	8	9	9	8
9	7	3	6	4
6	5	4	9	4
3	4	4	7	12
7	7	4	7	7
9	8	4	2	4
7	4	5	5	8
10	7	7	5	8
10	7	8	6	4
6	6	5	6	6
7	7	3	7	7
4	9	9	8	9
7	7	8	8	6
3	7	10	8	5
2	5	7	12	11
7	6	4	8	12
4	7	9	10	7
3	5	6	7	4
9	3	7	9	9

Aqui recorreu-se ao modelo estabelecido no Exemplo 3 (NPA≤1/36 → Soma=2; 1/36<NPA≤3/36 → Soma=3; 3/36<NPA≤6/36 → Soma=4; ...;15/36<NPA≤21/36 → Soma=7;....; 33/36<NPA≤35/36 → Soma=11; 35/36<NPA≤1 → Soma=12).

Exemplo

Tem levantado bastante polémica o seguinte exemplo de decisão estratégica:

Num concurso é dada a escolher ao concorrente uma de 3 portas. Atrás de uma delas está um carro e atrás de cada uma das outras duas está uma ovelha. O concorrente escolhe uma das portas (sem a abrir) e o apresentador, que sabe exactamente qual é a porta que esconde o carro, abre, de entre as duas portas que restam, uma onde está uma ovelha. Nesse momento pergunta ao concorrente se deseja ou não trocar a porta que escolheu pela outra porta que ainda está fechada. O primeiro pensamento que ocorre é que não há qualquer vantagem em trocar pois temos agora apenas duas portas e o carro tanto pode estar atrás de uma como da outra. No entanto, se se calcular a probabilidade do concorrente ganhar o carro, trocando de porta, verifica-se que esta é igual a 2/3 (basta fazer a árvore de probabilidades para a sequência de experiências "escolha" seguida de "troca" e considerar como acontecimentos relevantes "ganhar o carro" e "ganhar ovelha"). Para os mais reticentes uma simulação talvez os faça reconsiderar a sua posição inicial. Não há qualquer dúvida de que ao escolher uma porta ao acaso a probabilidade de ela esconder o carro é igual a 1/3. Para simular o decorrer de 100 destes concursos vamos então considerar que o concorrente escolheu a boa porta sempre que o valor do NPA estiver entre 0 e 1/3. Nestes casos, quando ele trocar de porta, ficará com a "ovelha" mas, em compensação, ficará com o carro em todos os outros casos (se ele tiver escolhido inicialmente a "ovelha", a porta que resta terá obrigatoriamente o carro pois o apresentador encarregou-se de eliminar a outra porta que também tinha "ovelha"!...). Eis o resultado da simulação obtida a partir da nossa lista de NPA's:

NPA	O que ganha não trocando	O que ganha trocando	NPA	O que ganha não trocando	O que ganha trocando	NPA	O que ganha não trocando	O que ganha trocando
0,84205	Ovelha	Carro	0,40632	Ovelha	Carro	0,848744	Ovelha	Carro
0,965131	Ovelha	Carro	0,676239	Ovelha	Carro	0,722927	Ovelha	Carro
0,761648	Ovelha	Carro	0,552387	Ovelha	Carro	0,079614	Carro	Ovelha
0,359825	Ovelha	Carro	0,20842	Carro	Ovelha	0,09815	Carro	Ovelha
0,054705	Carro	Ovelha	0,102768	Carro	Ovelha	0,147229	Carro	Ovelha
0,466613	Ovelha	Carro	0,493374	Ovelha	Carro	0,150888	Carro	Ovelha
0,8143	Ovelha	Carro	0,638416	Ovelha	Carro	0,086141	Carro	Ovelha
0,449515	Ovelha	Carro	0,090759	Carro	Ovelha	0,19746	Carro	Ovelha
0,901502	Ovelha	Carro	0,552418	Ovelha	Carro	0,466389	Ovelha	Carro
0,862762	Ovelha	Carro	0,507097	Ovelha	Carro	0,613583	Ovelha	Carro
0,395195	Ovelha	Carro	0,415666	Ovelha	Carro	0,210044	Carro	Ovelha
0,420519	Ovelha	Carro	0,469764	Ovelha	Carro	0,053714	Carro	Ovelha
0,124664	Carro	Ovelha	0,765629	Ovelha	Carro	0,737348	Ovelha	Carro
0,537707	Ovelha	Carro	0,451921	Ovelha	Carro	0,702749	Ovelha	Carro
0,033277	Carro	Ovelha	0,523063	Ovelha	Carro	0,908485	Ovelha	Carro
0,024371	Carro	Ovelha	0,213326	Carro	Ovelha	0,442821	Ovelha	Carro
0,558313	Ovelha	Carro	0,283191	Carro	Ovelha	0,153907	Carro	Ovelha

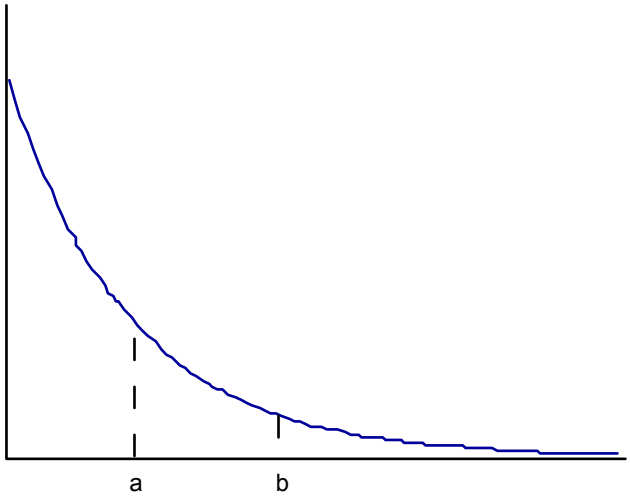
0,087859	Carro	Ovelha	0,429387	Ovelha	Carro	0,735276	Ovelha	Carro
0,069915	Carro	Ovelha	0,221549	Carro	Ovelha	0,358037	Ovelha	Carro
0,774156	Ovelha	Carro	0,039495	Carro	Ovelha	0,490216	Ovelha	Carro
0,810469	Ovelha	Carro	0,708764	Ovelha	Carro	0,71323	Ovelha	Carro
0,825587	Ovelha	Carro	0,983754	Ovelha	Carro	0,623757	Ovelha	Carro
0,2983	Carro	Ovelha	0,655705	Ovelha	Carro	0,129629	Carro	Ovelha
0,818893	Ovelha	Carro	0,89068	Ovelha	Carro	0,302539	Carro	Ovelha
0,55792	Ovelha	Carro	0,578713	Ovelha	Carro	0,444822	Ovelha	Carro
0,540352	Ovelha	Carro	0,755072	Ovelha	Carro	0,806147	Ovelha	Carro
0,00784	Carro	Ovelha	0,789583	Ovelha	Carro	0,377823	Ovelha	Carro
0,209145	Carro	Ovelha	0,702971	Ovelha	Carro	0,19629	Carro	Ovelha
0,221584	Carro	Ovelha	0,087455	Carro	Ovelha	0,970551	Ovelha	Carro
0,389183	Ovelha	Carro	0,103532	Carro	Ovelha	0,99576	Ovelha	Carro
0,379011	Ovelha	Carro	0,996667	Ovelha	Carro	0,569285	Ovelha	Carro
0,478208	Ovelha	Carro	0,480287	Ovelha	Carro	0,161851	Carro	Ovelha
0,696311	Ovelha	Carro	0,109918	Carro	Ovelha	0,753139	Ovelha	Carro
0,683382	Ovelha	Carro						

Como se verifica, nas 100 realizações simuladas deste concurso o concorrente ganharia o carro em 67 dessas realizações se decidisse por trocar de porta!...

Modelo Exponencial

O modelo Exponencial é muito utilizado na modelação do tempo de vida de equipamentos e na modelação do tempo entre chegadas de clientes em sistemas de filas de espera. A sua função de densidade de probabilidade é $f(x)=c\ \exp(-cx)$, para $x\geq 0$, $f(x)=0$, para $x<0$. Prova-se que o seu valor médio é igual a $1/c$ e que a área determinada por um intervalo $[a,b]$, entre o gráfico de f e o eixo dos xx, pode calculada através da expressão $\exp(-ca) - \exp(-cb)$.

Suponhamos então que este é um modelo razoável para descrever os intervalos de tempo ente chegadas sucessivas de carros a uma portagem. Se soubermos que o tempo médio entre essas chegadas é de 2 minutos então bastará fazer $c=0.5$ no modelo. Com esta informação podemos calcular a probabilidade de que entre a chegada de dois carros decorra menos do que 3 minutos ($\exp(-0.5x0) - \exp(-0.5x3)$), mais do que 5 minutos ($\exp(-0.5x5)$), ou qualquer outra que nos interesse.



Podemos até fazer uma simulação dos intervalos de tempo entre chegadas sucessivas. A metodologia é bastante simples: sabemos que a área entre o ponto 0 e um ponto a qualquer (com a>0) é sempre um número entre 0 e 1. Reciprocamente, dado um número entre 0 e 1, existe um e um só ponto a a que corresponde uma área entre 0 e a igual a esse número. Como os NPA's, acima referidos, tomam sempre valores no intervalo [0,1], a simulação de uma observação do modelo exponencial consiste então em resolver em ordem a a a equação NPA=área entre 0 e a. Mais precisamente, NPA=1-exp(-ca), ou seja, a= -1/c ln(1-NPA). Na tabela seguinte está uma exemplificação deste procedimento para c=0.5 e para os primeiros 20 NPA's. Acrescentámos ainda uma coluna com os intantes de chegada a partir do instante t=0.

NPA	-2 ln(1-NPA)	Instantes de Chegada
0,84205	3,690954	3m 42s
0,965131	6,712314	10m 24s
0,761648	2,868013	13m 16s
0,359825	0,892027	14m 10s
0,054705	0,112516	14m 16s
0,466613	1,257016	15m 32s
0,8143	3,367246	18m 54s
0,449515	1,193911	20m 02s
0,901502	4,635438	24m 43s
0,862762	3,972077	28m 42s
0,395195	1,005698	29m 42s
0,420519	1,091245	30m 48s
0,124664	0,266295	31m 04s
0,537707	1,543113	32m 36s
0,033277	0,067687	32m 40s
0,024371	0,049346	32m 43s
0,558313	1,634308	34m 22s
0,087859	0,183921	34m 32s
0,069915	0,144959	34m 41s
0,774156	2,975822	37m 39s

Modelo Poisson e Modelo Geométrico

Qualquer deles tem por suporte o conjunto {0,1,2,3, ...} e uma função massa de probabilidade dada por uma expressão algébrica, que envolve um único parâmetro. Mais precisamente, a cada elemento k do suporte, o **modelo Poisson** atribui a probabilidade

$$p_k = e^{-\lambda}\,\frac{\lambda^k}{k!}$$

Consoante o valor de λ seja maior ou menor assim a zona do suporte de maiores valores para a probabilidade se desloca para a direita ou para a esquerda. O valor médio e a variância do modelo de Poisson são iguais a λ. Muitos dos livros de estatística trazem tabelas deste modelo para diversos valores de λ e o software Excel também possui uma função interna que permite calcular as probabilidades referentes ao modelo Poisson para qualquer valor de λ.

Quanto ao **modelo Geométrico**, a sua função massa de probabilidade é dada por

$$p_k = p \times (1-p)^k$$

e é utilizado, principalmente, para se calcular a probabilidade de se ter k insucessos antes de ter um primeiro sucesso. O parâmetro p corresponde à probabilidade de sucesso. Por exemplo, ao jogar um dado, a probabilidade de só sair um 6 no décimo lançamento é

$$\frac{1}{6} \times \left(1 - \frac{1}{6}\right)^9$$

que se obtém substituindo o parâmetro p por 1/6 (probabilidade de sucesso) e o k por 9 (número de insucessos antes do primeiro sucesso).

3.8. Modelo Normal

Objectivos a atingir:

- ✓ Salientar a importância deste modelo referindo o Teorema Limite Central.
- ✓ Referir as principais características de um modelo Normal ou Gaussiano.
- ✓ Calcular probabilidades com base nesta família de modelos recorrendo ao uso de uma tabela da função de distribuição de uma Normal Standard.

O modelo Normal é um dos modelos mais utilizados em Estatística, devendo a sua relevância a um dos teoremas mais importantes da teoria da Probabilidade – o Teorema do Limite Central. Efectivamente, como veremos no módulo da Inferência Estatística, este teorema é a base de técnicas de inferência estatística largamente utilizadas, ao descrever as distribuições de amostragem, para a média e a proporção, como sendo aproximadamente normais.

O Teorema do Limite Central (TLC) pode ser enunciado brevemente da seguinte forma: *Qualquer característica aleatória que possa ser encarada como uma soma de muitas outras características aleatórias independentes, com variância finita, tem uma distribuição que se aproxima da*

distribuição Normal. Essa aproximação é tanto melhor quanto maior for o número de parcelas da soma. Muitas características de interesse ligadas a fenómenos naturais (altura de um indivíduo, perímetro do tronco de uma árvore, peso de um certo tipo de fruto, etc) podem ser encaradas como resultantes do contributo (de forma aditiva) de muitas variáveis. O TLC justifica a utilização do modelo Normal na modelação deste tipo de grandezas.

Começa-se por apresentar o modelo Normal, como um modelo com suporte em R e cuja função densidade tem uma forma característica, fazendo lembrar a forma de um sino. Quando falamos no modelo Normal, falamos mais propriamente numa família de distribuições caracterizadas por dois parâmetros: o valor médio e o desvio padrão. Devem ser apresentados vários exemplos de distribuições, com o mesmo parâmetro valor médio e desvio padrão diferente, assim como várias distribuições com valor médio diferente e o mesmo desvio padrão. Uma característica importante deste modelo, a ser utilizada posteriormente no cálculo das probabilidades, é a simetria da função densidade relativamente a um eixo vertical com abcissa igual ao valor médio.

Deve seguidamente ser ensinado aos alunos, como com base numa tabela existente para o modelo Normal standard, isto é, de valor médio igual a 0 e desvio padrão igual a 1, se pode calcular um valor aproximado para a probabilidade de a variável assumir valores de um intervalo.

Posteriormente ensinar-se-á a técnica da estandardização, isto é, o processo de transformar qualquer modelo Normal de valor médio μ e desvio padrão σ, $N(\mu,\sigma)$, no modelo standard $N(0,1)$, de forma a permitir a utilização das tabelas existentes para este modelo para calcular a probabilidade de qualquer variável, com distribuição Normal, assumir valores de um qualquer intervalo.

Chamar a atenção para a propriedade dos valores de uma variável com distribuição Normal de valor médio μ e desvio padrão σ estarem com probabilidade aproximadamente igual a 1, concentrados num intervalo de amplitude 6σ e centrado no valor médio. Chamar a atenção para a regra dos 68 – 95 – 100, apresentada no módulo da Estatística e que agora pode ser formalmente demonstrada.

4. Introdução à Inferência Estatística

4.1. Parâmetro e estatística. Distribuição de amostragem.

Objectivos a atingir:

- ✓ Apresentar as ideias básicas de um tipo de raciocínio com que os alunos são confrontados pela primeira vez, em que a partir das propriedades estudadas num conjunto de dados, se procurarão tirar conclusões para um conjunto de dados mais vasto.

Neste módulo deve-se começar por recordar o que foi estudado no capítulo da produção e aquisição de dados, objecto de estudos estatísticos. Deve ser recordado que nos processos utilizados para produzir dados, foi realçada a necessidade de que estes devem ser baseados em métodos probabilísticos. Neste contexto destacam-se os métodos de amostragem que conduzem às *amostras aleatórias*, em que existe um mecanismo aleatório que faz com que um elemento da população faça parte da amostra, assim como *as experimentações controladas*, em que cada indivíduo é escolhido aleatoriamente para lhe ser atribuído um tratamento. As razões invocadas na altura prendem-se sobretudo com a recolha de amostras não enviesadas.

Neste módulo compreender-se-á todo o alcance desta necessidade de aleatoriezar o processo de recolha de dados, pois veremos que esse facto nos vai permitir utilizar a teoria das probabilidades para descrever o comportamento do processo associado com a recolha e sumariação dos dados, um grande número de vezes.

Um dos objectivos que se tem ao recolher uma amostra de uma População que se pretende estudar é o de retirar conclusões sobre os *parâmetros* (características numéricas) dessa População. Assim, quando se pretende estimar (obter um valor aproximado) um determinado parâmetro, considera-se uma função conveniente que só dependa dos elementos da amostra – *estatística.*

Deve-se chamar a atenção para o facto de se utilizar um tipo de raciocínio indutivo, em que se vai procurar tirar conclusões, indo do particular para o geral. Este tipo de raciocínio é contrário ao tipo de raciocínio matemático, essencialmente dedutivo.

Exemplos

1. Diga se são verdadeiras ou falsas as afirmações seguintes:
Se uma amostra é aleatória qualquer elemento da população tem a mesma probabilidade de pertencer à amostra.
Resposta: A afirmação é verdadeira se se tratar de uma amostra aleatória simples. Em outros processos aleatórios de amostragem, nomeadamente o da recolha de uma amostra estratificada, nem todos os elementos da população têm igual probabilidade de pertencerem à amostra.
Uma estatística é um número que se calcula a partir da amostra.
Resposta: Verdadeira.
Os parâmetros utilizam-se para estimar estatísticas.
Resposta: É falso, pois são as estatísticas que se utilizam para estimar parâmetros.
A média populacional é um parâmetro.

Resposta: A afirmação é verdadeira. Efectivamente a média populacional ou valor médio é um parâmetro que se estima a partir da média calculada a partir de uma amostra.

Reduz-se o enviesamento, aumentando a dimensão da amostra.

Resposta: Não é verdade. O enviesamento não tem nada a ver com a dimensão da amostra, mas sim com o processo de selecção da amostra. Reduz-se o enviesamento recorrendo a processos aleatórios de recolha da amostra.

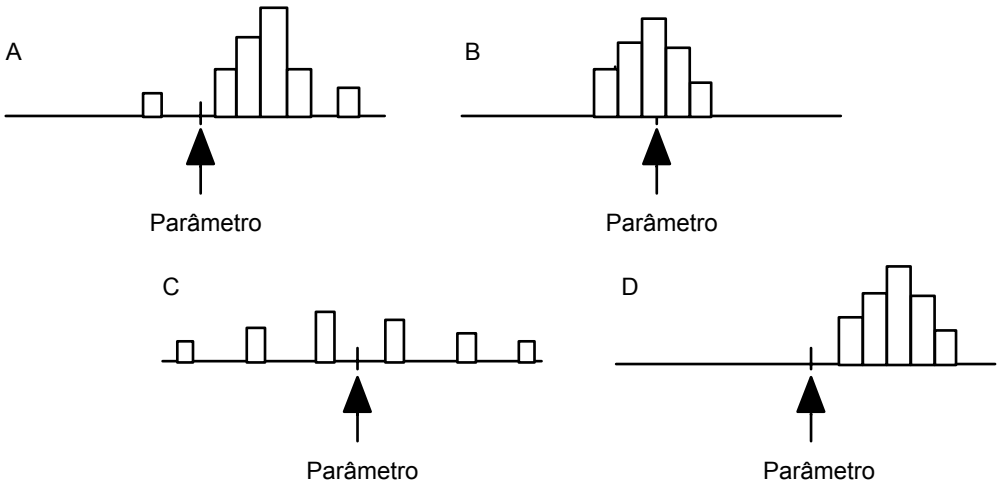
Reduz-se a variabilidade da distribuição de amostragem da média, aumentando a dimensão da amostra.

Resposta: É verdade. Quanto maior for a dimensão da amostra a partir da qual se calcula a estatística, menor é a variabilidade presente na distribuição de amostragem.

Um estimador com uma distribuição de amostragem com uma pequena variabilidade fornece necessariamente boas estimativas.

Resposta: Não é verdade. Só fornecerá boas estimativas se o estimador não for enviesado, isto é se a média da distribuição de amostragem coincidir com o valor do parâmetro a estimar.

2. Suponha que dispõe de 4 estatísticas A, B, C e D para estimar um parâmetro. As distribuições de amostragem (para amostras da mesma dimensão) das 4 estatísticas têm o seguinte aspecto:



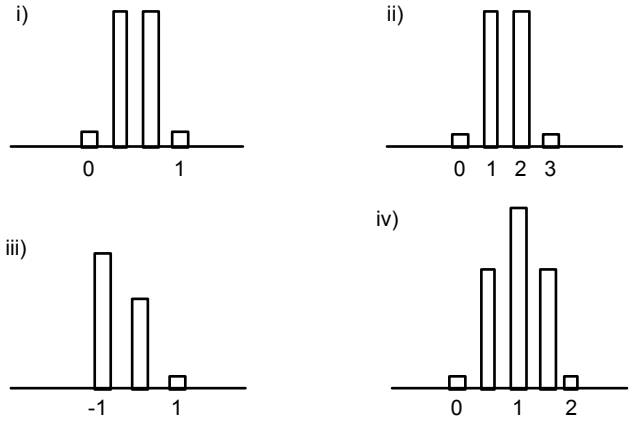
Qual das estatísticas escolheria?

Resposta: A estatística B, pois é a que apresenta o centro da distribuição de amostragem a coincidir com o valor do parâmetro. A estatística A apresenta enviesamento, do mesmo modo que a D, e além disso apresenta maior variabilidade. A estatística C embora não apresente enviesamento, apresenta uma grande variabilidade.

3. Considere os seguintes processos e as estatísticas associadas (Adaptado de Cryer et al, 1997):

Processo	Estatística
1) Lança 3 moedas	Nº de caras obtidas
2) Lança 3 moedas	Média do nº de caras obtidas
3) Lança 4 moedas	Nº de caras/2
4) Extrai 2 nºs do conjunto (-1, -1, -1, 2)	Média dos valores obtidos

Qual das seguintes representações gráficas descreve a distribuição de amostragem da estatística considerada?



Resposta: (1, II), (2, I), (3, IV), (4, III)

4. Considere um grupo de 4 alunos com pesos de 16, 17, 18 e 19 quilogramas, respectivamente.

a) Quantas amostras diferentes, de dimensão 2, pode extrair da população constituída pelos 4 alunos? Considere a extracção sem reposição e com reposição.

b) Considere a extracção com reposição. Qual a probabilidade de obter duas vezes o mesmo elemento na amostra?

c) Se a sua população tivesse 5 elementos, e extraísse amostras de dimensão 2, com reposição, qual a probabilidade de obter amostras com o elemento repetido? E se a sua população tivesse dimensão 10000? O que pode concluir relativamente à extracção com reposição e sem reposição, se a dimensão da população for muito grande?

Resposta: a)

Extracção sem reposição	Extracção com reposição
(16, 17), (16, 18), (16, 19)	(16, 16), (16, 17), (16, 18), (16, 19)
(17, 16), (17, 18), (17, 19)	(17, 16), (17, 17), (17, 18), (17, 19)
(18, 16), (18, 17), (18, 19)	(18, 16), (18, 17), (18, 18), (18, 19)
(19, 16), (19, 17), (19, 18)	(19, 16), (19, 17), (19, 18), (19, 19)
num total de 12 amostras	num total de 16 amostras

Como temos 4 amostras com os elementos iguais, em 16 amostras, a probabilidade é 4/16 = 1/4.

Se a população tivesse 5 elementos a probabilidade de extrairmos amostras de dimensão 2, com elementos iguais seria 1/5. Se a população tivesse 10000 elementos aquela probabilidade seria 1/10000.

Se a dimensão da população for muito grande, a probabilidade de extrairmos elementos iguais é extremamente pequena, pelo que nas populações de grande dimensão, os processos de extracção com reposição e sem reposição são praticamente equivalentes.

5. Considere uma população constituída pelos elementos 1, 2, 3, 4 e 5. Pretende estimar o valor médio desta população, pelo que decide recolher uma amostra de dimensão 2, e calcular a sua

média. Obtenha a distribuição de amostragem do estimador utilizado para estimar o valor médio da população.

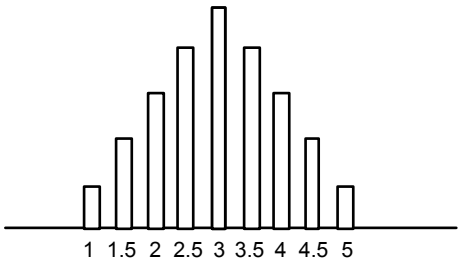
Resolução: A metodologia seguida para obter a distribuição de amostragem consiste em obter todas as amostras de dimensão 2, calcular o valor da estatística média para cada uma delas e depois representar a distribuição dos valores obtidos.

População (1, 2, 3, 4, 5)

Amostras	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)
		(2,1)	(2,2)	(2,3)	(2,4)	(3,4)	(4,4)	(5,4)	
			(3,1)	(3,2)	(3,3)	(4,3)	(5,3)		
				(4,1)	(4,2)	(5,2)			
					(5,1)				
média	1	1.5	2	2.5	3	3.5	4	4.5	5

De acordo com a tabela anterior obtemos a seguinte distribuição de amostragem para o estimador média

média	1	1.5	2	2.5	3	3.5	4	4.5	5
probabilidade	1/25	2/25	3/25	4/25	5/25	4/25	3/25	2/25	1/25



Características da distribuição de amostragem:

Valor médio = 3

Desvio padrão = 1

Algumas observações:

O centro da distribuição de amostragem do estimador média utilizado para estimar o valor médio da população (igual a 3), coincide com o parâmetro a estimar .

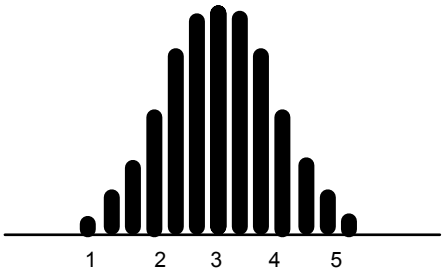
O desvio padrão da população inicial é igual a $\sqrt{2}$, enquanto que o desvio padrão da média, calculada a partir de amostras de dimensão 2 é 1 ($\sqrt{2}$ / $\sqrt{2}$ =1 – resultado a explicar posteriormente).

6. Repita o processo do exemplo anterior, mas considerando agora amostras de dimensão 3.

Resolução: Utiliza-se a mesma metodologia seguida no processo anterior, mas agora considerando amostras de dimensão 3, o que torna o problema mais trabalhoso. Abstemo-

nos de descrever todas as amostras possíveis, em número de 5^3=125, e apresentamos a distribuição de amostragem da média

média	1	1.33	1.67	2	2.33	2.67	3	3.33	3.67	4	4.33	4.67	5
Proba.	.008	.024	.048	.080	.120	.144	.152	.144	.120	.080	.048	.024	.008



Características da distribuição de amostragem:

Valor médio = 3

Desvio padrão = 0.816

Algumas observações:

O centro da distribuição de amostragem do estimador média utilizado para estimar o valor médio da população (igual a 3), coincide com o parâmetro a estimar .

O desvio padrão da população inicial é igual a $\sqrt{2}$, enquanto que o desvio padrão da média, calculada a partir de amostras de dimensão 3 é 0.816 ($\sqrt{2}/\sqrt{3}$ =0.816 – resultado a explicar posteriormente).

A variabilidade apresentada pela distribuição de amostragem é inferior à obtida quando se consideram amostras de dimensão 2. *Este resultado indicia que quanto maior for a dimensão da amostra, menor é a variabilidade apresentada pela distribuição de amostragem*.

7. No Departamento de Estatística há 5 docentes que são professores associados, dos quais 3 são mulheres – Maria, Ana, Rita e 2 são homens – Pedro e Tiago. Se representarmos por **p** a percentagem de homens que são professores associados, temos que p=2/5. Suponhamos que pretendíamos estimar esta proporção utilizando a proporção $\hat{\mathbf{p}}$ de homens em amostras de dimensão 2. Então vamos construir todas as amostras desta dimensão para obter a distribuição de amostragem da estatística utilizada:

Amostra	$\hat{\mathbf{p}}$	Amostra	$\hat{\mathbf{p}}$
Maria, Maria	0	Rita, Pedro	1/2
Maria, Ana	0	Rita, Tiago	1/2
Maria, Rita	0	Pedro, Maria	1/2
MariaPedro	1/2	Pedro, Ana	1/2
MariaTiago	1/2	Pedro, Rita	1/2
Ana, Maria	0	Pedro, Pedro	2/2
Ana, Ana	0	Pedro, Tiago	2/2
Ana, Rita	0	Tiago, Maria	1/2

Ana, Pedro	1/2	Tiago, Ana	1/2
Ana, Tiago	1/2	Tiago, Rita	1/2
Rita, Maria	0	Tiago, Pedro	2/2
Rita, Ana	0	Tiago, Tiago	2/2
Rita, Rita	0		

A partir da tabela anterior é possível obter a distribuição de amostragem da estatística $\hat{\mathbf{p}}$:

$\hat{\mathbf{p}}$	0	.5	1
Probabilidade	9/25	12/25	4/25

$E(\hat{\mathbf{p}}) = 2/5$ e $Var(\hat{\mathbf{p}}) = 3/25$

Repare-se que o valor médio da estatística $\hat{\mathbf{p}}$ coincide com o valor do parâmetro **p** que se está a estimar.

4.2. Noção de estimativa pontual. Estimação de um valor médio e de uma proporção. Distribuição de amostragem. Construção de estimativas intervalares ou intervalos de confiança para o valor médio e para a proporção.

Objectivos a atingir:

- ✓ Apresentar as ideias básicas de um processo de inferência estatística, em que se usam estatísticas para tomar decisões acerca de parâmetros.
- ✓ Mostrar toda a potencialidade da Estatística, que nos permite tirar conclusões e tomar decisões, indo do particular para o geral, quantificando o erro cometido nessa tomada de decisões.

À estatística utilizada para estimar um determinado parâmetro chamamos *estimador* do parâmetro. Quando se recolhe uma amostra, calcula-se a partir dos dados da amostra recolhida o valor do estimador, que dá uma *estimativa* do parâmetro. Se se recolher outra amostra da mesma População e da mesma dimensão, é natural obter uma estimativa para o parâmetro, diferente da primeira. Quantas amostras recolhermos, quantas as estimativas diferentes que podemos obter para o parâmetro. É importante chamar a atenção para que não podemos dizer qual *das estimativas pontuais* é melhor, já que não se conhece o valor do parâmetro a estimar.

Esta *variabilidade* apresentada pelas estimativas, é inerente à aleatoriedade da escolha da amostra e uma questão que se coloca é a de saber se o estimador que se está a considerar é um "bom" estimador ou não, isto é, se por um lado as estimativas que produz são próximas umas das outras, ou apresentam uma grande variabilidade, e se por outro lado, no caso de apresentarem pequena variabilidade, se serão aproximadas do parâmetro que se pretende estimar.

A resposta a esta questão é dada construindo a distribuição de todos os valores apresentados pela *estatística* que se está a utilizar para estimar o parâmetro, para todas as amostras possíveis, da mesma dimensão. A esta distribuição dá-se o nome de *distribuição de amostragem da estatística*. Ao aleatoriezar o processo de selecção das amostras, faz com que se possa utilizar a distribuição de amostragem de uma estatística para descrever o comportamento dessa estatística, quando se usa para estimar um determinado parâmetro. Se a média da distribuição de amostragem da estatística coincidir com o valor do parâmetro a estimar, dizemos que o estimador é *não enviesado*. Quanto à variabilidade apresentada pela distribuição de amostragem da estatística, quanto menor ela for, mais perto do parâmetro estão as estimativas obtidas a partir da estatística considerada.

A compreensão das diferenças entre *parâmetro* e *estatística* e do que é uma *distribuição de amostragem*, é a base dos processos de Inferência Estatística.

Os parâmetros que se procurarão estimar são:
o *valor médio* – medida de localização do centro da distribuição dos valores assumidos por uma dada variável, cujo estimador será a *média* de uma amostra de observações dessa variável;
a *proporção* ou frequência relativa com que se verifica uma determinada característica na População, cujo estimador será a *proporção* de vezes que essa característica se verifica nos elementos da amostra recolhida dessa População.

Sendo a noção de distribuição de amostragem a base da maior parte das técnicas de inferência estatística, é importante exemplificar o seu processo de construção, podendo para começar, considerar um dos casos mais simples que é o de estimar um *valor médio*.

Nesta altura deve-se também chamar a atenção e exemplificar o papel desempenhado pela dimensão da amostra, para a precisão dos resultados, na medida em que diminui a variabilidade apresentada pela distribuição de amostragem.

Começa-se aqui a introduzir o conceito de *confiança estatística*, como resultado do estudo da distribuição de amostragem.

Uma vez trabalhado e entendido o conceito de distribuição de amostragem, deve-se recordar um resultado teórico, já enunciado no módulo da Probabilidade, com a maior relevância para a Estatística, conhecido pelo *Teorema do Limite Central*. Este teorema legitima, de certa maneira, a grande utilização do modelo Normal como modelo de variáveis que resultem de medições de grandezas naturais como a altura, peso, etc, que se admitem serem o resultado de um grande número de contribuições cumulativas. Estando a média e a proporção neste caso, este resultado poupa o trabalho de estar a obter as suas distribuições de amostragem, desde que as amostras tenham dimensão suficientemente grande, e o processo utilizado para as recolher tenha sido aleatório.

O processo da construção de distribuições de amostragem estende-se à proporção amostral, estatística utilizada para estimar o parâmetro *proporção* (probabilidade) de elementos da População que verificam uma determinada propriedade. O processo a seguir para o estudo da proporção pode ser o de considerar esta como um caso particular de uma média quando os elementos que têm a propriedade em estudo são representados por 1, enquanto que os outros são representados por 0.

Finalmente introduzir-se-á o conceito de *intervalo de confiança* tanto para o valor médio da característica em estudo da População, como para a proporção com que uma determinada característica está presente nos elementos da População.

Deverá ser chamada a atenção para a interpretação correcta do que é que se entende por *confiança*, ao considerar um intervalo de confiança.

Considera-se importante que os alunos interpretem a *amplitude* do intervalo, como a maior ou menor precisão, isto é, como a *margem de erro* dos resultados obtidos quando se considera uma determinada confiança e uma determinada dimensão para a amostra. Deverá ser realçado o facto de a amplitude do intervalo de confiança depender da variabilidade da estatística utilizada.

O conceito de intervalo de confiança deverá ser trabalhado de forma a que os alunos fiquem aptos a interpretar resultados veiculados pela comunicação social tais como: “o resultado da sondagem é de 76% com uma margem de erro de ± 3 pontos percentuais”.

Os exemplos relacionados com as sondagens em tempo de campanhas eleitorais ou relativamente a outros problemas têm muito interesse, pois muito facilmente se encontram exemplos na comunicação social. Aliás, deve ser incentivada a leitura dos jornais e a recolha de assuntos que enunciem resultados objecto de tratamento estatístico.

Deverão também ser trabalhados vários exemplos que permitam descobrir o efeito de se utilizarem amostras de maior ou menor dimensão na determinação dos intervalos de confiança, quando a dimensão da População é muito superior à dimensão das amostras com que se trabalha. Sugere-se que se apresente a seguinte regra: *Se a dimensão da População for muito superior à dimensão da amostra (por exemplo 100 vezes superior), a variabilidade da distribuição de amostragem é a mesma para qualquer dimensão da População*. Esta regra traduz uma característica importante dos processos de amostragem, na medida em que traduz o facto de as distribuições de amostragem não dependerem (muito) da dimensão da População.

Finalmente deve-se chamar a atenção para o facto de que se as amostras recolhidas forem enviesadas, os intervalos de confiança também virão enviesados, não tendo portanto qualquer utilidade.

Exemplo

No exemplo seguinte, continuamos a estudar alguns conceitos já abordados nos exemplos anteriores, mas agora num contexto de uma situação mais elaborada e mais perto de uma situação real.

Para exemplificar a diferença entre *parâmetro* e *estatística*, assim como o que se entende por *distribuição de amostragem* de uma estatística, conceito fundamental em Inferência Estatística, vamos apresentar uma População finita, isto é vamos considerar um conjunto de indivíduos com algumas características comuns, algumas das quais nos interessam estudar, nomeadamente a variável Altura e a propriedade de cada indivíduo ser ou não do sexo masculino.

Considere a seguinte tabela onde se apresentam os 97 trabalhadores de uma determinada empresa:

Número	Nome	Estado civil	Idade	Altura	Nº filhos
1	Alexandra Almeida	solteira	26	160	0
2	Alexandre Carmo	casado	30	174	2
3	Alda Morais	casada	37	160	3
4	Ana Ribeiro	casada	23	159	1
5	Ana Cristina Santos	casada	26	156	2
6	Ana Cristina Oliveira	solteira	25	153	0
7	Anabela Pais	divorciada	33	156	3
8	António Couto	solteiro	24	177	0
9	António Fernandes	casado	42	161	5
10	António Pinto	casado	51	171	1
11	Armando Ferreira	casado	48	167	1
12	Carlos Matos	casado	37	165	1
13	Carlos Sampaio	casado	40	174	2
14	Cristina Vicente	casada	39	160	2
15	Cristina Zita	casada	27	164	1
16	Dora Ferreira	casada	50	170	4
17	Elsa Sampaio	casada	45	160	4
18	Fernando Barroso	casado	43	164	3
19	Fernando Martins	casado	29	165	1
20	Fernando Santos	divorciado	32	174	2
21	Filomena Silva	solteira	20	165	0
22	Francisco Gomes	casado	26	174	0
23	Isabel Soares	solteira	22	156	0
24	Isabel Silva	casada	34	148	2
25	João Morais	casado	44	171	2
26	João Sousa	solteiro	25	176	0
27	Luis Horta	casado	35	169	2
28	Luis Sousa	casado	37	170	0
29	Luis Ribeiro	casado	49	170	1
30	Manuel Santos	casado	54	175	4
31	Manuel Pereira	divorciado	47	162	3
32	Manuel Teixeira	casado	50	173	2
33	Margarida Almeida	casada	51	166	1
34	Margarida Simões	casada	47	161	4
35	M. Adelina Azevedo	solteira	25	148	0
36	M. Alexandra Almeida	solteira	26	158	0
37	M. Alexandra Ribeiro	casada	39	157	3

38	M. Cristina Carvalho	casada	41	158	2
39	M. Cristina Freire	divorciada	38	161	1
40	M. de Fátima Osório	casada	33	164	1
41	M. Fernanda Rocha	solteira	29	154	0
42	M. Isabel Frade	casada	38	164	2
43	M. Isabel Santos	solteira	26	164	0
44	M. Luisa Faria	casada	35	164	2
45	M. Manuel Trindade	casada	29	167	0
46	M. Manuela Lino	casada	33	159	3
47	M. Nazaré Pinto	solteira	29	162	0
48	M. Neusa Lopes	casada	34	163	2
49	M. Olga Martins	casada	27	165	0
50	M. Paula Pitarra	casada	29	160	3
51	M. Paula Garcês	solteira	25	150	0
52	M. Rosário Gomes	solteira	27	155	0
53	M. Rute Costa	solteira	45	160	0
54	M. Rute Rita	solteira	23	165	0
55	M. Teresa António	casada	46	147	2
56	M. Teresa Bento	casada	54	158	1
57	M. Teresa Garcia	solteira	22	154	0
58	Mário Martins	casado	29	171	1
59	Mário Reis	casado	43	172	0
60	Nuno Simões	casado	43	176	2
61	Nuno Ventura	solteiro	28	175	0
62	Olga Martins	solteira	29	159	0
63	Oscar Trigo	casado	35	169	1
64	Osvaldo	casado	44	172	1
65	Paulo Nunes	casado	38	169	1
66	Paulo Martins	solteiro	41	173	1
67	Paulo Santos	solteiro	51	172	1
68	Paulo Valente	casado	45	168	2
69	Pedro Casanova	casado	46	175	1
70	Pedro Dalo	casado	37	166	1
71	Pedro Martins	casado	39	174	2
72	Pedro Lisboa	casado	44	163	2
73	Pedro Sintra	solteiro	40	170	0
74	Pedro Valente	casado	32	161	0
75	Pedro Viriato	casado	26	169	0
76	Rita Amaral	solteira	23	165	0
77	Rita Bendito	solteira	29	159	0
78	Rita Évora	casada	34	162	1
79	Rita Seguro	solteira	30	163	0
80	Rita Valente	casada	35	170	2
81	Rufo Almeida	solteiro	29	171	0
82	Rui André	solteiro	31	165	0
83	Rui Martins	casado	34	167	0
84	Rui Teixeira	casado	44	166	2
85	Rui Vasco	casado	45	178	2
86	Sérgio Teixeira	divorciado	40	174	2
87	Sílvio Lino	divorciado	44	161	0
88	Tânia Lopes	casada	27	160	0
89	Tânia Martins	solteira	25	162	0
90	Teresa Adão	casada	26	163	1
91	Teresa Paulo	solteira	28	164	0
92	Teresa Vasco	casada	30	157	0
93	Vera Mónica	solteira	25	161	0
94	Vera Patrícia	solteira	26	154	0
95	Vera Teixeira	casada	31	162	1
96	Vitor Santos	casado	37	173	2
97	Vitor Zinc	solteiro	49	169	0

No que diz respeito às variáveis Sexo, Idade, Altura e Número de filhos a população anterior tem as seguintes características:

Tabela 1

Sexo	Freq. abs.	Freq. rel.
Feminino	52	0.536
Masculino	45	0.464

	97	1.000

Tabela 2

Variável	Valor Médio	Desvio padrão	Mínimo	Máximo
Idade	35.19	8.84	20	54
Altura	164.57	7.05	147	178
Nº filhos	1.13	1.21	0	5

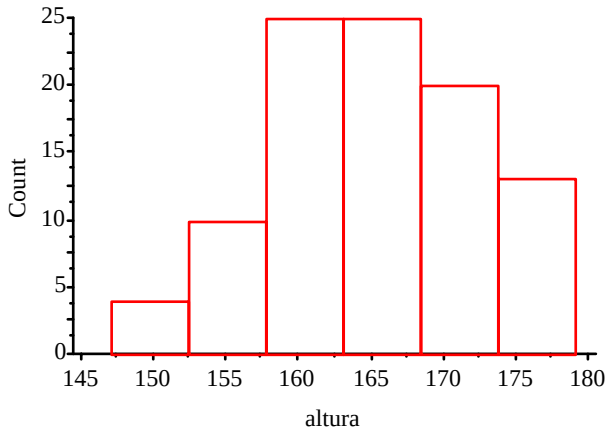
Repare-se que para a variável Sexo não calculámos nem a média nem o desvio padrão, já que se trata de uma variável qualitativa.

Distribuição da variável Altura

Para construir o histograma para a variável Altura utilizámos um programa de Estatística, chamado Statview, que automaticamente constitui as classes depois de dizermos com quantas classes queremos construir o histograma, mas também se pode utilizar o Excel. Neste caso teremos de começar por dar os limites para as intervalos de classe que se querem considerar.

Tabela 3

Bar:	From: (□)	To: (<)	Count:	Percent:
1	147	152.333	4	4.124%
2	152.333	157.667	10	10.309%
3	157.667	163	25	25.773%
4	163	168.333	25	25.773%
5	168.333	173.667	20	20.619%
6	173.667	179	13	13.402%



Exemplo 1 – Estimação da altura

Para exemplificar o processo de estimação de uma característica da População, nomeadamente a variável Altura, pretendemos seleccionar uma amostra aleatória, pelo que apresentamos uma

tabela de números aleatórios, onde a primeira coluna referencia o número de linha para facilidade de consulta da tabela:

Tabela 4

1	013	581	704	400	988	100	938	997	298	856
2	623	023	137	118	929	567	939	964	963	752
3	490	083	021	121	378	551	866	913	807	504
4	339	358	318	108	069	677	437	740	568	911
5	788	770	497	267	700	869	369	114	836	241
6	492	291	887	676	412	898	843	850	656	196
7	893	761	037	810	468	719	324	854	469	783
8	537	160	210	070	665	264	100	820	073	287
8	605	648	400	391	511	860	203	953	036	272
10	153	115	795	410	046	868	179	512	423	321
11	164	239	068	327	070	488	181	099	333	237
12	489	988	790	798	093	081	523	410	319	759
13	790	565	366	895	084	982	020	822	827	618
14	226	750	758	647	791	774	529	789	008	138
15	549	919	473	901	594	338	884	673	235	631
16	094	570	597	509	211	043	490	543	018	747
17	439	199	498	092	644	079	740	644	408	765
18	012	029	055	194	288	490	873	945	993	761
19	119	362	265	419	603	912	506	347	898	686
20	504	497	702	447	912	581	371	138	357	863

Vamos, por exemplo, considerar a linha 6 para começar a seleccionar aleatoriamente 15 trabalhadores da empresa, identificados pelo número, pelo que seleccionamos 15 números de 2 algarismos:

49 22 91 88 76 76 41 28 (98) 84 38 50

65 61 96 89 37

No processo anterior tivemos de seleccionar 16 números, uma vez que o um deles não correspondia a nenhum número de trabalhador. Os trabalhadores seleccionados foram os seguintes:

Tabela 5

	Nome	Est. civil	Idade	Altura	Nº filhos	Sexo
49	M. Olga Martins	casada	27	165	0	0
22	Francisco Gomes	casado	26	174	0	1
91	Teresa Paulo	solteira	28	164	0	0
88	Tânia Lopes	casada	27	160	0	0
76	Rita Amaral	solteira	23	165	0	0
76	Rita Amaral	solteira	23	165	0	0
41	M. Fernanda Rocha	solteira	29	154	0	0
28	Luis Sousa	casado	37	170	0	1
84	Rui Teixeira	casado	44	166	2	1
38	M. Cristina Carvalho	casada	41	158	2	0
50	M. Paula Pitarra	casada	29	160	3	0
65	Paulo Nunes	casado	38	169	1	1
61	Nuno Ventura	solteiro	28	175	0	1
96	Vitor Santos	casado	37	173	2	1
89	Tânia Martins	solteira	25	162	0	0

Mais à frente faremos algumas considerações sobre o processo de selecção da amostra.

As tabelas seguintes apresentam algumas características da amostra, nomeadamente a proporção de elementos do sexo masculino, a média das idades, a média das alturas e a média do número de filhos.

Tabela 6

Sexo	Freq. abs.	Freq. rel.
Feminino	9	0.600
Masculino	6	0.400
	15	1.000

Tabela 7

Variável	Média	Desvio padrão	Mínimo	Máximo
Idade	30.80	6.74	23	44
Altura	165.33	6.07	154	175
Nº filhos	0.67	1.05	0	3

O valor 165.33 obtido para a média da amostra constituída pelas alturas, diz-se que é uma estimativa do parâmetro “altura média” da População de onde a amostra foi retirada, assim como o valor 0.4 é uma estimativa para o parâmetro “proporção de pessoas do sexo masculino” na População. Quando comparamos os valores das *estatísticas* com os valores dos *parâmetros* (neste caso conhecidos), verificamos que obtemos valores relativamente próximos.

Observação: Recordamos que a situação que estamos a tratar, em que se conhecem os parâmetros das características em estudo, não é a situação corrente. De um modo geral as Populações têm dimensão muito grande ou mesmo infinita, de forma que o seu estudo só pode ser feito mediante a recolha de uma amostra.

Variabilidade das estatísticas

Se recolhermos outras amostras da mesma dimensão não vamos obter os mesmos valores como estimativas dos parâmetros. Vamos repetir o processo de seleccionar 50 amostras de dimensão 15 da população em estudo, e registar na tabela seguinte a média das alturas obtidas e a proporção de pessoas do sexo masculino, para cada uma das amostras seleccionadas:

Tabela 8

Amostra	Média	Prop.H	Amostra	Média	Prop.H	Amostra	Média	Prop.H	Amostra	Média	Prop.H
1	164.27	0.47	14	165.27	0.53	27	163.53	0.33	40	164.40	0.47
2	162.67	0.33	15	164.40	0.40	28	164.60	0.60	41	163.20	0.33
3	165.27	0.53	16	161.00	0.27	29	162.80	0.33	42	162.80	0.20
4	165.67	0.60	17	162.07	0.33	30	164.40	0.47	43	165.07	0.40
5	164.47	0.33	18	163.60	0.53	31	164.20	0.53	44	164.93	0.47
6	163.93	0.40	19	162.47	0.33	32	162.87	0.40	45	163.93	0.40
7	162.07	0.20	20	162.87	0.40	33	163.87	0.47	46	165.27	0.47
8	163.67	0.27	21	161.13	0.40	34	165.53	0.47	47	161.67	0.33
9	166.33	0.73	22	164.07	0.47	35	165.93	0.53	48	163.87	0.47
10	163.93	0.60	23	163.67	0.47	36	164.80	0.40	49	166.20	0.60
11	163.40	0.60	24	167.60	0.67	37	166.40	0.53	50	165.40	0.40
12	162.67	0.40	25	165.33	0.53	38	164.53	0.47			
13	162.87	0.40	26	164.53	0.33	39	163.73	0.53			

Como se verifica da tabela anterior os valores das estatísticas média e proporção, *estimativas pontuais* dos respectivos parâmetros “altura média” e “proporção de indivíduos do sexo masculino na População”, calculadas a partir das diferentes amostras, **variam** de amostra para amostra.

Então, se considerarmos duas amostras de alturas e a partir dos seus elementos obtivéssemos duas estimativas pontuais para o valor do parâmetro “altura média”, se não se conhecesse o seu valor, como é que poderíamos saber qual das estimativas era melhor, isto é, estava mais perto do valor do parâmetro? Na realidade não temos maneira de escolher entre os dois valores, qual o melhor. Podemos, no entanto, ir estudar como se comportam as diferentes estimativas, para ver se se retira algum padrão da sua distribuição.

Consideremos então em primeiro lugar a amostra de dimensão 50, constituída pelas médias das 50 amostras de alturas e depois a amostra, também de dimensão 50, constituída pelas proporções de indivíduos do sexo masculino existentes nas 50 amostras.

Estudo da amostra das médias das amostras de alturas

Amostras de dimensão 15

a) Características amostrais

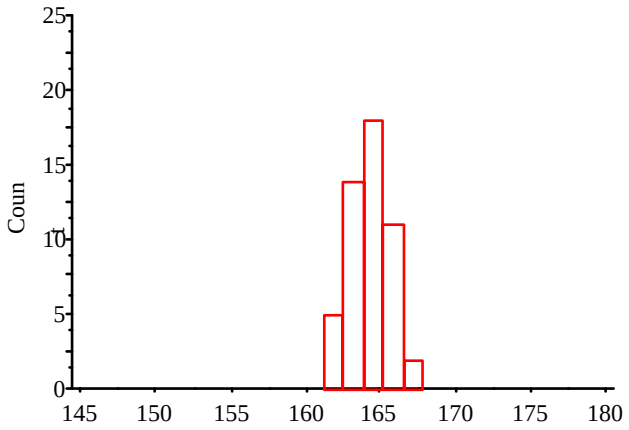
Tabela 9

Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:
164.063	1.393	.197	1.939	.849	50
Minimum:	**Maximum:**	**Range:**	**Sum:**	**Sum of Sqr.:**	**# Missing:**
161	167.6	6.6	8203.16	1345931.707	0
# < 10th %:	**10th %:**	**25th %:**	**50th %:**	**75th %:**	**90th %:**
5	162.27	162.87	164	165.07	165.8
# > 90th %:					
5					

b) Histograma

Tabela 10

Bar:	From: (□)	To: (<)	Count:	Percent:	
1	161	162.34	5	10%	
2	162.34	163.68	14	28%	
3	163.68	165.02	18	36%	-Mode
4	165.02	166.36	11	22%	
5	166.36	167.7	2	4%	



Considerámos o histograma com a mesma escala que o considerado para a variável Altura, para ser mais fácil de retirarmos as seguintes conclusões, que também eram evidentes a partir da tabela das características amostrais:

a) A distribuição da amostra das médias apresenta uma variabilidade muito pequena;

b) A distribuição da amostra das médias faz-se de forma aproximadamente simétrica em torno do valor 164.1, que é um valor muito próximo da altura média da característica Altura;

c) Da tabela das características amostrais verificamos que 80% dos elementos da amostra das médias estão no intervalo [162.27, 165.8], de amplitude 3.53, que contém o valor do parâmetro “altura média”.

Por outro lado se em vez de termos seleccionado amostras de dimensão 15, tivéssemos seleccionado amostras de dimensão 30, o que é que viria diferente?

Amostras de dimensão 30

Repetimos então a experiência de seleccionar 50 amostras de dimensão 30, da característica Altura, calculámos a média de cada uma das amostras e considerámos a amostra constituída pelas 50 médias. De seguida apresentamos essa amostra e o estudo descritivo dessa amostra (utilizámos o programa STATVIEW):

Tabela 11

Amostra	Média	Prop.H	Amostra	Média	Prop.H	Amostra	Média	Prop.H	Amostra	Média	Prop.H
1	164.83	0.43	14	165.83	0.53	27	166.50	0.50	40	165.40	0.40
2	163.93	0.43	15	164.60	0.40	28	166.00	0.63	41	164.10	0.50
3	164.57	0.50	16	163.47	0.40	29	163.90	0.40	42	162.23	0.33
4	163.67	0.47	17	163.80	0.40	30	166.63	0.50	43	165.33	0.50
5	164.87	0.40	18	163.73	0.50	31	164.47	0.57	44	165.27	0.50
6	163.43	0.37	19	165.67	0.53	32	164.37	0.43	45	164.20	0.37
7	163.50	0.33	20	163.03	0.43	33	164.97	0.47	46	164.13	0.37
8	164.37	0.37	21	162.77	0.43	34	161.80	0.33	47	165.47	0.50
9	164.57	0.47	22	163.43	0.40	35	164.33	0.37	48	163.90	0.47
10	163.07	0.40	23	164.13	0.53	36	162.67	0.33	49	164.47	0.40
11	163.90	0.53	24	165.57	0.50	37	166.90	0.50	50	163.33	0.50
12	163.20	0.47	25	164.93	0.43	38	165.53	0.57			
13	163.70	0.43	26	161.87	0.30	39	163.13	0.40			

a) Características amostrais

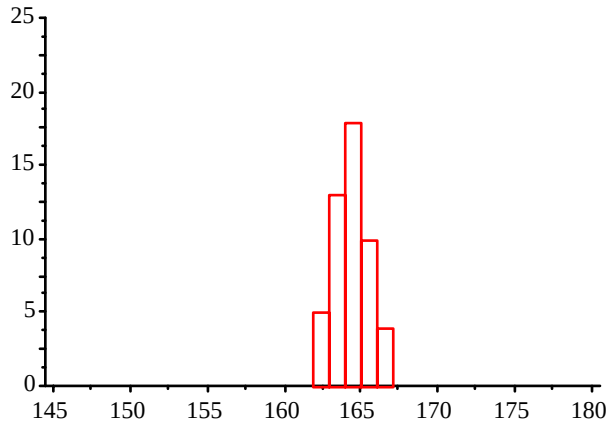
Tabela 12

Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:
164.269	1.165	.165	1.358	.709	50
Minimum:	**Maximum:**	**Range:**	**Sum:**	**Sum of Sqr.:**	**# Missing:**
161.8	166.9	5.1	8213.47	1349288.324	0
# < 10th %:	**10th %:**	**25th %:**	**50th %:**	**75th %:**	**90th %:**
5	162.9	163.47	164.165	164.97	165.75
# > 90th %:					
5					

b) Histograma

Tabela 13

Bar:	From: (□)	To: (<)	Count:	Percent:	
1	161.8	162.84	5	10%	
2	162.84	163.88	13	26%	
3	163.88	164.92	18	36%	-Mode
4	164.92	165.96	10	20%	
5	165.96	167	4	8%	



Comparando os resultados das duas experiências, verificamos o seguinte:

a) Do mesmo modo que para as amostras de dimensão 15, o histograma correspondente à amostra das médias das amostras de dimensão 30, é aproximadamente simétrico;
b) A média da amostra é 164.26, valor este que é mais próximo do valor do parâmetro “altura média” (164.57), do que quando consideramos as amostras de dimensão 15;
c) A variabilidade apresentada pela amostra das médias de amostras de dimensão 30 é inferior à apresentada pelas amostras de dimensão 15;
d) Da tabela das características amostrais verificamos que 80% dos elementos da amostra das médias estão no intervalo [162.9, 165.75], de amplitude 2.85, que contém o valor do parâmetro “altura média”, enquanto que no caso das amostras de dimensão 15, este intervalo tinha amplitude 3.53.

Os resultados anteriores levam-nos a pensar que quanto maior for a dimensão das amostras consideradas menor será a variabilidade entre as médias dessas amostras. Então se recolhêssemos uma amostra de dimensão 40, da característica Altura, esperaríamos que a média dessa amostra desse uma estimativa mais perto do parâmetro “altura média”, do que a média de uma amostra de dimensão 15 ou 30.

Quando recolhemos as 50 amostras e calculámos a média de cada uma dessas amostras, ficámos com uma ideia do comportamento da *estatística* média, que resumimos no seguinte:

- ◊ Quanto maior for a dimensão da amostra, espera-se que seja melhor a estimativa fornecida pela *estatística* "média" para o *parâmetro* "valor médio" da característica que se está a estudar;
- ◊ Quando consideramos amostras da mesma dimensão, a média varia de amostra para amostra, mas apresenta um comportamento característico, de uma distribuição aproximadamente simétrica, com pequena variabilidade.

Alguns resultados que ainda podemos obter a partir da consulta da tabela que apresenta os elementos da amostra são: A percentagem de elementos da amostra que estão nos seguintes intervalos é:

Tabela 14

Intervalo	**Nºelem.**	**%**
[média –desvio padrão, média + desvio padrão] [163.104, 165.434]	34	68%
[média –2xdesvio padrão, média + 2xdesvio padrão] [161.939, 166.599]	46	92%
[média –3xdesvio padrão, média + 3xdesvio padrão] [160.774, 167.764]	50	100%

Assim, os intervalos anteriores podem ser encarados como uma "espécie" de *intervalos de confiança* para o parâmetro que estamos a estimar, na medida em que, por exemplo, confiamos que 100% das estimativas calculadas a partir de várias amostras estarão no intervalo [160.774, 167.764], etc.

E se em vez de 50 amostras considerássemos todas as amostras possíveis (diferentes) que se podem extrair da População? No nosso caso, se quiséssemos amostras de dimensão 30, teríamos de seleccionar 97^{30} amostras! Isto seria muito trabalhoso, mas só assim é que teríamos verdadeiramente a distribuição da média das amostras de dimensão 30, isto é, os diferentes valores que a variável

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_{30}}{30}$$

pode assumir e a probabilidade de assumir esses valores, a que chamamos *distribuição de amostragem* da média. Estamos a representar a variável que está a ser estudada –por um X, pelo que X_1 representa a 1ª vez que se foi seleccionar um elemento, X_2 representa a a 2ª vez que se foi seleccionar um elemento, etc.

Observação 1: Repare-se que a média $\bar{X}$ é uma variável aleatória pois os seus valores dependem dos valores das variáveis X_1, X_2, …, X_{30}. Quando observamos um valor de X_1, que representamos por x_1, um valor de X_2, que representamos por x_2, etc, e substituímos esses

valores observados na expressão da média, obremos um valor observado para a média, que representamos por $\bar{x}$.

Observação 2: Aproveitamos para chamar a atenção para o facto de a amostragem ser feita com reposição, pois cada vez que se selecciona um elemento ele é reposto, antes de seleccionar o seguinte. Esta observação é sobretudo relevante para Populações de dimensão pequena (como a considerada no nosso estudo), em que a composição da População sofre alteração quando se retiram alguns elementos, o que não sucede com Populações de grande dimensão (que é normalmente a situação de interesse em Estatística).

Distribuição de amostragem da média

Então para obter a distribuição de amostragem da média teremos de considerar todas as amostras possíveis e depois calcular as respectivas médias?

Felizmente não é necessário estar com tanto trabalho, graças a um dos resultados mais importantes das Probabilidades, conhecido como o Teorema do Limite Central e que nos fornece um modelo matemático para a distribuição de amostragem da média:

Teorema do Limite Central – Suponhamos que se recolhe uma amostra de dimensão n de uma população muito grande X, com valor médio μ e desvio padrão σ. Então, se a dimensão da amostra for suficientemente grande ($n \geq 30$) a distribuição de amostragem da média pode ser aproximada por uma distribuição Normal com valor médio μ e desvio padrão $\sigma/\sqrt{n}$.

Exemplo 1 (continuação) - Tendo em consideração o Teorema do Limite Central, como espera que seja a distribuição de amostragem da média, para amostras de dimensão 30?

Espera-se que possa ser aproximadamente modelada por um modelo de uma Normal com valor médio 164.57 e desvio padrão $7.05/\sqrt{30} = 1.287$.

Calcule:

a) Um valor aproximado para a probabilidade de a média diferir do valor médio de uma quantidade inferior a 5 décimas
   Resolução: Pretende-se

$$P\left(\left|\overline{X}\right| - \mu \leq 0.5\right)$$

$$P\left(-0.5 \leq \overline{X} - \mu \leq 0.5\right) = P\left(\frac{-0.5}{1.287} \leq \frac{\overline{X} - \mu}{1.287} \leq \frac{0.5}{1.287}\right) \cong \phi(0.3885) - (-0.3885) \cong 0.30$$

$$\cong \Phi(0.3885) - \Phi(-0.3885) \cong 0.30$$

b) Um valor aproximado para a probabilidade de a média estar no intervalo [161.94, 166.60]
   Resolução: Pretende-se

$$P(161.94 \le \overline{X} \le 166.60)$$

$$P(161.94 \le \overline{X} \le 166.60) = P\left(\frac{161.94-164.57}{1.287} \le \frac{\overline{X}-164.57}{1.287} \le \frac{166.60-164.57}{1.287}\right) \cong$$

$$\cong \phi(1.58) - \phi(-2.04) \cong 0.92$$

Observação: Repare-se no valor obtido para esta probabilidade e para o valor obtido para a proporção obtida na página anterior para a proporção correspondente a um intervalo análogo.

c) Calcule o valor de z tal que $P(-z \le \frac{\overline{X}-164.57}{1.287} \le z) \cong 0.95$

$$P(-z \le \frac{\overline{X}-164.57}{1.287} \le z) \cong \phi(z) - \phi(-z) \cong 0.95$$

$2\,\phi(z) - 1 \cong 0.95$ implica que z= 1.96

Agora a probabilidade anterior pode-se escrever

$$P(\overline{X}) - 1.96 \times 1.287 \le 164.57 \le \overline{X} + 1.96 \times 1.287 = .95$$

e o intervalo

$$[\overline{X} - 1.96 \times 1.287\ ,\ \overline{X} + 1.96 \times 1.287]$$

diz-se que é um *intervalo com uma confiança* de 95% para o parâmetro "valor médio da Altura ou Altura média".

Intervalo de confiança para o valor médio da Altura

Anteriormente exemplificámos a construção do que chamamos um intervalo de confiança, com uma confiança de 95%. Como é que se interpreta esta confiança? O que é que significa?

Considere a primeira amostra que recolheu, de dimensão 30, cuja média deu 164.83. Se substituir este valor na expressão considerada anteriormente obtém o intervalo [162.31, 167.35]. Se considerar a segunda amostra cujo média deu 163.93, obtém o intervalo [161.41, 166.45]. Se finalmente considerar as 50 amostras recolhidas e as respectivas médias, obtém os seguintes 50 intervalos

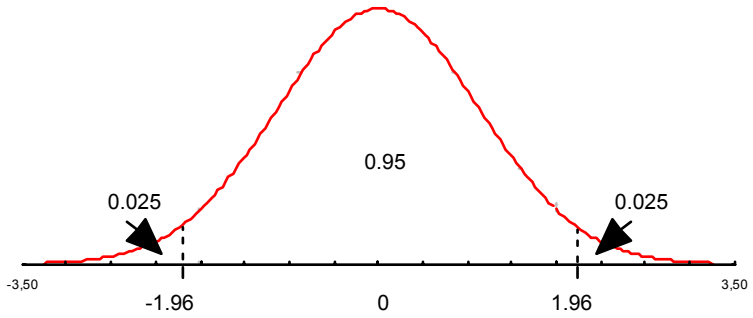
[162.31, 167.36]	[161.38, 166.42]	[160.24, 165.29]	[161.94, 166.99]	[161.58, 166.62]
[161.41, 166.46]	[160.68, 165.72]	[160.91, 165.96]	[161.84, 166.89]	[159.71, 164.76]
[162.04, 167.09]	[161.18, 166.22]	[161.61, 166.66]	[162.44, 167.49]	[162.81, 167.86]
[161.14, 166.19]	[163.31, 168.36]	[163.04, 168.09]	**[159.28, 164.32]**	[160.74, 165.79]
[162.34, 167.39]	[162.08, 167.12]	[162.41, 167.46]	[161.81, 166.86]	[161.68, 166.72]
[160.91, 165.96]	[160.94, 165.99]	**[159.34, 164.39]**	[160.14, 165.19]	[161.61, 166.66]
[160.98, 166.02]	[161.28, 166.32]	[163.98, 169.02]	[164.38, 169.42]	[162.94, 167.99]
[161.84, 166.89]	[161.21, 166.26]	[163.48, 168.52]	[163.01, 168.06]	[161.38, 166.42]
[162.04, 167.09]	[160.14, 165.19]	[161.38, 166.42]	[160.61, 165.66]	[161.94, 166.99]
[160.54, 165.59]	[160.51, 165.56]	[160.11, 165.16]	[162.88, 167.92]	[160.81, 165.86]

Destes 50 intervalos, verifica-se que 48 contêm o valor do parâmetro “Altura média”, que é 164.57, enquanto que 2 – assinalados a escuro , não o contêm. Quando falamos em 95% de confiança, significa que se considerássemos 100 intervalos, esperaríamos que aproximadamente 95 contivessem o valor do parâmetro e 5 não o contivessem.

Como ao fazer um estudo sobre um parâmetro desconhecido, só se recolhe uma amostra, temos *confiança* que a que recolhemos seja uma das “boas”, que vai dar origem a um intervalo que contenha o valor desse parâmetro.

Suponhamos que pretendíamos determinar um intervalo de confiança para o valor médio μ de uma variável aleatória, desconhecido, e admitamos também que não conhecíamos o desvio padrão σ dessa variável. Recolhemos uma amostra de dimensão n, suficientemente grande (≥30) e calculamos a média. Pelo Teorema do Limite Central sabemos que a distribuição das diferentes médias, obtidas a partir das diferentes amostras que se podem recolher de dimensão n, pode ser aproximada por uma distribuição Normal de valor médio μ e desvio padrão $\sigma/\sqrt{n}$. Mas se o σ é desconhecido como é que fazemos? Substituímos o σ pelo desvio padrão s da amostra considerada, pois ainda continuamos a ter uma distribuição aproximadamente Normal. Então, se pretendermos o valor de z tal que

$$P(-z \leq \frac{\overline{X} - \mu}{S/\sqrt{n}} \leq z) \cong 0.95$$



obtemos o valor z=1.96, pelo que a expressão para o intervalo de confiança com uma confiança de 95% é a seguinte

$$[\bar{X} - 1.96 \times S/\sqrt{n} \ , \ \bar{X} + 1.96 \times S/\sqrt{n}]$$

Para uma dada amostra que se recolhe calcula-se a média e o desvio padrão, substitui-se na expressão anterior e obtém-se o intervalo pretendido, que se chama também uma estimativa intervalar para o parâmetro.

O que é que significa aumentar a confiança?

Suponhamos que pretendíamos uma confiança de 99%. Então, neste caso pretendíamos o valor de z tal que

$$P(-z \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq z) \cong 0.99$$

Por um processo análogo ao que nos permitiu obter o valor 1.96 para z, quando a probabilidade considerada era 0.95, agora obteríamos para z o valor 2.596. Então, a forma do intervalo de confiança, com uma confiança de 99% será

$$[\bar{X} - 2.596 \times S/\sqrt{n} \ , \ \bar{X} + 2.596 \times S/\sqrt{n}]$$

Se utilizando a mesma amostra construirmos um intervalo com 95% de confiança e outro com 99% de confiança, a amplitude deste último é superior à do primeiro, como se verifica da expressão dos intervalos. Assim, aumentar a confiança implica aumentar a amplitude do intervalo, pelo que aumenta a nossa confiança em que um qualquer intervalo que se construa contenha o valor do parâmetro que estamos a estimar.

Exemplo 2 – Estimação da proporção

Suponhamos agora que se pretende estimar uma proporção, nomeadamente o *parâmetro* "proporção de indivíduos do sexo masculino" na População objecto de estudo. Neste caso o valor do parâmetro também é conhecido, situação que normalmente não se verifica, como já referimos, e daí a razão de se recolher uma amostra para o estimar.

Consideremos então a primeira amostra de dimensão 15, recolhida no exemplo anterior: a proporção de indivíduos do sexo masculino presentes na amostra é 0.4, que é uma estimativa para o parâmetro "proporção de indivíduos do sexo masculino" na População.

Antes de prosseguirmos com o estudo da proporção, vejamos um pouco mais em detalhe como é que se pode calcular uma proporção.

Uma proporção é uma média?

Suponhamos que estamos a estudar uma População quanto à presença ou ausência de uma determinada propriedade, isto é, cada indivíduo da População tem ou não tem essa propriedade. Admitimos que essa propriedade se verifica na População com uma probabilidade p (normalmente desconhecida). Se ao observar o indivíduo verificarmos que tem a propriedade, anotamos um 1, enquanto que se verificarmos que não tem a propriedade anotamos um 0. Então podemos representar a População, quanto a essa propriedade por uma variável X, que pode assumir o valor 1 ou 0, respectivamente com probabilidade p (probabilidade de ter a propriedade) ou (1-p) (probabilidade de não ter a propriedade).

Como é que se pode interpretar p? Repare-se que p é a frequência relativa com que o 1 se verifica na População relativamente à propriedade em estudo, e não é mais do que a média do conjunto constituído pelos 0's e 1's.

Analogamente quando recolhemos uma amostra, constituída por 1's e 0's conforme os elementos observados tenham ou não tenham a propriedade, a média desta amostra dá-nos a proporção (amostral) de 1's, ou seja, uma estimativa pontual para a proporção (populacional) ou probabilidade com que a propriedade em estudo se verifica na População.

Do que acabamos de referir, depreende-se que o estudo do parâmetro p "proporção de indivíduos da população que verificam determinada propriedade" se reduz ao estudo do parâmetro "valor médio de uma população representada por 1's e 0's, conforme a propriedade está ou não presente nos indivíduos da população".

Suponhamos então que no caso da População em estudo, se pretende estudar a probabilidade p de um indivíduo seleccionado ao acaso, ser do sexo masculino, a partir da proporção de indivíduos do sexo masculino presentes numa amostra de dimensão 30:

De acordo com a notação introduzida anteriormente, esta estimativa tem que ser um valor da variável

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_{30}}{30}$$

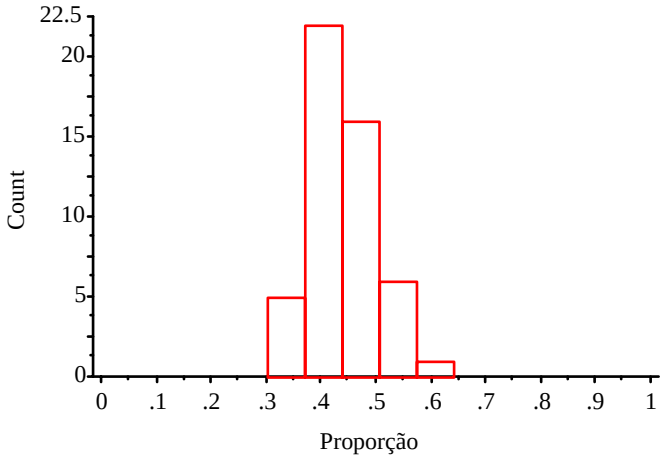
em que cada uma das variáveis X_1, X_2, … , X_n é idêntica à variável X, considerada anteriormente, que só pode tomar os valores 1 e 0, com probabilidades, respectivamente p e (1-p).
Como esta variável X tem valor médio p (probabilidade de um indivíduo ser do sexo masculino) e desvio padrão $\sqrt{p(1-p)}$ vem, tendo em conta o Teorema do Limite Central, que o estimador da probabilidade p, usualmente representado por $\hat{p}$, tem uma distribuição de amostragem que pode ser aproximada pelo modelo Normal, com valor médio 0.464 (valor do parâmetro p) e desvio padrão $\sqrt{\frac{0.464\ (1-0.464)}{30}} = 0.091$.

Para as 50 amostras de dimensão 30, consideradas no exemplo anterior, os valores das proporções de indivíduos do sexo masculino estão apresentadas na tabela 11. O estudo da distribuição desses valores apresenta-se a seguir:

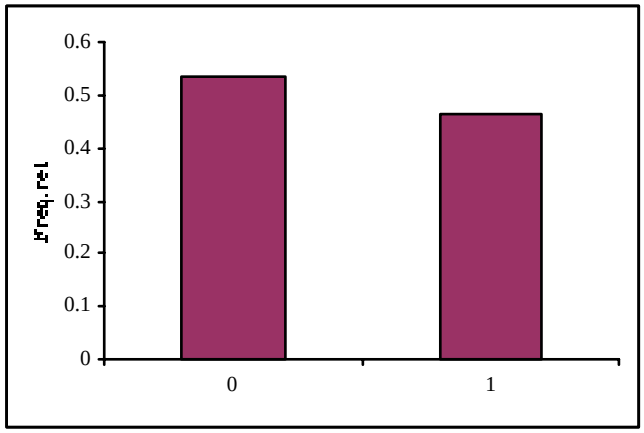
Características amostrais

Mean:	Std. Dev.:	Std. Error:	Variance:	Coef. Var.:	Count:
.444	.072	.01	.005	16.221	50
Minimum:	**Maximum:**	**Range:**	**Sum:**	**Sum of Sqr.:**	**# Missing:**
.3	.63	.33	22.22	10.129	0
# < 10th %:	**10th %:**	**25th %:**	**50th %:**	**75th %:**	**90th %:**
5	.35	.4	.43	.5	.53
# > 90th %:					
3					

Histograma



É interessante verificar que a variável X em estudo tem a seguinte distribuição



Como se verifica é uma distribuição discreta que assume os valores 0 e 1, enquanto que para a distribuição de amostragem da média de variáveis com distribuição idêntica à de X, já se começa a sugerir o padrão da Normal (independentemente do resultado teórico dado pelo Teorema do Limite Central que justifica essa aproximação).

Como resultado das observações anteriores podemos enunciar o seguinte resultado, para a distribuição de amostragem da proporção:

Suponhamos que se recolhe uma amostra de dimensão n de uma população muito grande X, em que cada elemento da população tem uma determinada propriedade ou não a tem. Seja **p** a proporção de elementos da população com essa propriedade. Então, se a dimensão da amostra for suficientemente grande (n≥30), a distribuição de amostragem da proporção $\hat{\mathbf{p}}$ pode ser aproximada por uma distribuição Normal com valor médio **p** e desvio padrão $\sqrt{p(1-p)}/\sqrt{n}$.

Observação: Para exemplificar a determinação da distribuição de amostragem da proporção, poderíamos ter seguido um caminho idêntico ao que fizemos para a média, sem termos considerado a proporção como um caso particular da média.

Intervalo de confiança para a proporção populacional p

Já que a proporção populacional p é um valor médio e a proporção amostral $\hat{p}$ é uma média, a expressão para o intervalo de confiança da proporção p deduz-se da que se obteve para o intervalo de confiança para o valor médio μ, fazendo as modificações adequadas:

		Onde está	Considera-se
	σ	$\sqrt{p(1-p)}$	
ou	s	$\sqrt{\hat{p}(1-\hat{p})}$	

Como o valor de p é desconhecido, a expressão para o intervalo de confiança, com uma confiança de 95% vem

$$[\hat{p} - 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$$

onde se substitui o valor do parâmetro desconhecido p, por uma estimativa $\hat{p}$.

O intervalo que se obtém para a primeira a mostra de dimensão 30 que se considerou é

[0.340, 0.520]

Efeito da dimensão da amostra na amplitude do intervalo de confiança

Vimos anteriormente que se se aumentar a confiança, então a amplitude do intervalo também aumenta, o que significa que podemos diminuir a amplitude, diminuindo a confiança. Sobre este ponto convém trabalhar com os alunos, alertando-os para o facto de que um intervalo com uma amplitude nula, isto é, com os limites inferior e superior iguais, tem confiança nula (é o que nós chamámos de estimativa pontual), enquanto que um intervalo com uma confiança de 100% tem uma amplitude infinita (que também não nos serve para nada!).

Da expressão do intervalo de confiança verifica-se que um outro processo de diminuir a amplitude do intervalo, é aumentar a dimensão da amostra. Sobre este ponto convém lembrar aos alunos o que se passa em dia de eleições, por exemplo para as legislativas, em que quando se fecham as urnas começam a sair os primeiros resultados sobre a forma de intervalos, para as percentagens previstas para cada um dos partidos. À medida que a noite vai avançando, que

corresponde a mais votos escrutinados, portanto a amostras de maior dimensão, os intervalos começam a diminuir de amplitude, até chegarmos ao valor correcto do parâmetros a estudar – percentagens obtidas diferentes partidos, quando todos os votos forem escrutinados.

Margem de erro de uma sondagem

É frequente vermos na comunicação social resultados de sondagens expressos na seguinte forma:
O resultado da sondagem é de 76%, com uma *margem de erro* de ± 3 pontos percentuais. A margem de erro não é mais do que metade da amplitude do intervalo de confiança de 95%, ou seja, se representarmos o intervalo de confiança na forma

$$\hat{p} \pm 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

a margem de erro é a parcela $1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, cujo valor é 0.03, quando se substitui a proporção por 0.76.
Assim, de acordo com o que já se viu anteriormente, para diminuir a margem de erro temos duas opções:

- ◊ Diminuir a confiança (não aconselhável)
- ◊ Aumentar a dimensão da amostra

Qual a dimensão da amostra necessária para se obter uma estimativa com uma determinada margem de erro?

Suponhamos que estamos interessados em estimar a proporção **p** de portugueses continentais, que são a favor das touradas, com touros de morte. Qual a dimensão da amostra necessária de forma a obter um intervalo de 95% de confiança com uma margem de erro de ± 3 pontos percentuais?

- ◊ Admitamos que resultados de sondagens anteriores nos permitem afirmar que a proporção que pretendemos estimar anda à volta de 5%.

  Da equação $1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.03$ vem que $n = \left(\frac{1.96}{0.03}\right)^2 \hat{p}(1-\hat{p})$, onde substituindo a estimativa da proporção por 0.05, se obtém para n o valor 203 (repare-se que n tem que ser inteiro).
- ◊ E se a estimativa de p fosse 0.50, em vez de 0.05? Qual a dimensão da amostra necessária? Neste caso obteríamos n = 10672.

◊ Calcule, na sua máquina de calcular ou no computador, o valor de $\left(\frac{1.96}{0.03}\right)^2 \hat{p}(1-\hat{p}$ para vários valores de $\hat{p}$ entre 0 e 1. Por exemplo considere valores a começar em 0.01 até 0.99, com passo igual a 0.01. O que é que conclui? Interprete os resultados obtidos na alínea a) e b), baseado no estudo que acabou de fazer.

◊ Se não tivermos uma ideia de um valor aproximado para o parâmetro que estamos a estimar, qual a dimensão para a amostra que se deve considerar?

Qual a influência da dimensão da População na determinação de um intervalo de confiança?

Suponha que pretende obter uma estimativa da proporção p de portugueses (continentais e insulares) que são a favor das touradas, com touros de morte. Qual a dimensão da amostra necessária de forma a obter um intervalo de 95% de confiança com uma margem de erro de ± 3 pontos percentuais? Admita que as estimativas para p, são as consideradas no exemplo anterior.
Os valores para as dimensões das amostras não vêm alterados, pelo que concluímos que quando a dimensão da População é muito grande, quando comparada com a dimensão das amostras consideradas, as dimensões das amostras necessárias são as mesmas, para obter a mesma margem de erro.

Tem sentido utilizar um intervalo de confiança para estimar um parâmetro quando a amostra coincide com a População?

Não tem sentido, pois na construção do intervalo de confiança entra-se com o valor observado da estatística, precisamente no pressuposto de que o valor do parâmetro é desconhecido.

Tem sentido construir um intervalo de confiança com base numa amostra enviesada?

Suponha que pretende estimar a proporção de professores do sexo feminino que leccionam o Ensino Secundário no ano lectivo de 2000/2001. Com base numa lista fornecida pelo Ministério da Educação, seleccionou aleatoriamente 120 professores, dos quais 83 são do sexo feminino.

◊ Construa um intervalo de 95% de confiança para a proporção de mulheres que leccionam no Ensino Secundário.

$\hat{p} = \frac{83}{120} = 0.69$

$$[0.69 - 1.96\times\sqrt{\frac{0.69\times(1-0.69)}{120}},\ 0.69 + 1.96\times\sqrt{\frac{0.69\times(1-0.69)}{120}}]$$

$$[0.69 – 0.08,\ 0.69 + 0.08]$$

Intervalo pretendido [0.61, 0.77]

A resposta também pode ser dada na seguinte forma: o resultado da sondagem é 69% com uma margem de erro de $\pm$ 8 pontos percentuais.

◊ Suponha que utiliza a amostra anterior para estimar a proporção de mulheres na população portuguesa. Este procedimento é correcto?
Não, porque a amostra é grandemente enviesada, pois a subpopulação dos professores tem uma proporção de mulheres largamente superior à população portuguesa. Assim, a estimativa de 69% obtida para a proporção de mulheres na amostra de professores, é uma estimativa muito enviesada da proporção de mulheres na população portuguesa.

De seguida apresentam-se algumas sugestões de exemplos a realizar pelos alunos, que podem servir para consolidar as assuntos abordados anteriormente:

Exemplo 1 – Qual a percentagem de alunos da escola que no último mês comprou pelo menos um livro, sem ser livro de estudo.

a) Pergunte junto de 80 alunos da escola, se compraram pelo menos um livro no último mês. Calcule a proporção de alunos que compraram pelo menos um livro no último mês. Este número é um parâmetro ou uma estatística?
b) Obtenha um intervalo de 90% de confiança para a proporção de alunos da escola que compraram pelo menos um livro no último mês.
c) Faça um pequeno relatório com a indicação do que representa o intervalo que obteve na alínea anterior.

Exemplo 2 – A Associação de Estudantes pretende obter uma estimativa de quantos alunos do 12º ano pretendem ir à viagem de finalistas.

a) Qual o parâmetro de interesse?
b) Pergunte a uma turma de alunos do 12º ano, quantos estão a prever ir à viagem de finalistas. Calcule a percentagem dos que pensam ir à viagem de finalistas.
c) Faça a mesma pergunta a outra turma de alunos do 12º ano e calcule a percentagem dos que prevêem ir à viagem de finalistas. O valor obtido para esta proporção foi o mesmo que o obtido anteriormente, na alínea b)? Comente os resultados obtidos.
d) Obtenha um intervalo de 95% de confiança para a proporção de alunos que pensam ir à viagem de finalistas. Qual a amostra que considerou? Explique porquê.

Exemplo 3 – Qual a lista para a Associação de estudantes que merece a preferência dos alunos?

a) Apresentam-se duas listas A e B, para a Associação de Estudantes. Investigue junto de 100 alunos em quem é que pensam votar nas próximas eleições e apresente os resultados numa tabela.
b) Calcule a proporção de alunos que pretendem votar na lista A.
c) Obtenha um intervalo de 95% de confiança para a proporção de alunos da escola que pensam votar na lista A. Qual a amplitude do intervalo obtido?
d) Se a amostra obtida tivesse 4 00 elementos, qual seria a amplitude do intervalo obtido anteriormente, se se tivesse obtido o mesmo valor para a proporção?

Exemplo 4 – Os estudantes estão satisfeitos com a política seguida para o ensino por este Governo?

Faça um estudo adequado que lhe permita responder a esta questão.

5. Bibliografia

COMAP (2000) – For all Practical Purposes. W. H. Freeman and Company. New York.

FREEDMAN, D. PISANI, R. PURVES, R., ADHIKARI, A. (1991) - *Statistics*. W. W. Norton & Company.

GRAÇA MARTINS, M. E. (1998) – *Introdução às Probabilidades e à Estatística*. Sociedade Portuguesa de Estatística.

GRAÇA MARTINS, M. E. , CERVEIRA, A. (1998) – *Introdução às Probabilidades e à Estatística*. Universidade Aberta.

IMAN, R. e CONOVER, W. (1983) - *A Modern Approach to Statistics*. John Wiley & Sons.

MANN, P. (1995) – *Introductory Statistics*. John Wiley & Sons.

MENDENHALL. W. BEAVER, R. (1994) – *Introduction to Probability and Statistics*. Duxbury Press.

MOORE, D. (1997) – *Statistics* – Concepts and Controversies. Freeman.

MOORE, D. (1995) – *The Basic Practice of Statistics,* Freeman.

MOORE, D., McCABE, G. (1993) – *Introduction to The Basic Practice of Statistics,* Freeman.

MURTEIRA, B. (1993) – *Análise Exploratória de dados – Estatística Descritiva*, McGraw-Hill de Portugal.

National Council of Teachers of Mathematics (1981) - *Teaching Statistics and Probability*, 1981 Yearbook, , Reston, EUA.

PARKS, H. et al. (1997) – *Mathematics in Life, Society & the World*, Prentice-Hall, Inc.

PARZEN, E. (1969) – *Modern Probability Theory and Its Applications*. New York.Wiley.

ROSSMAN, A. (1996) – *Workshop Statistics: discovery with data*. Springer-Verlag New York, Inc.

RUNYON, R. et al. (1996) – *Fundamentals of Behavioral Statistics*. McGraw-Hill Companies, Inc.

SIEGEL, A. (1988) – *Statistics and Data Analysis*. John Wiley & Sons.

TANNENBAUM, P. et al. (1998) – *Excursions in Modern Mathematics*. Prentice-Hall, Inc.

THIESSEN, H. (1997) – *Measuring the Real World*. John Wiley & Sons.